Title
Single-cell droplet microfluidics for metagenomics and cancer multiomics

Author
Demaree, Benjamin Robert

Publication Date
2020

Peer reviewed|Thesis/dissertation

Single-cell droplet microfluidics for metagenomics and cancer multiomics

by

Benjamin Demaree

DISSERTATION

Submitted in partial satisfaction of the requirements for degree of

DOCTOR OF PHILOSOPHY

in

Bioengineering

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO
AND
UNIVERSITY OF CALIFORNIA, BERKELEY

Approved:

DocuSigned by:
Adam Abate
E298E6C4A36E400...

Adam Abate
Chair

DocuSigned by:
Katherine Pollard
DocuSigned by:
[illegible]F1...

Katherine Pollard

Dorian Liepmann
2F6D3F3A132942E...

Committee Members

Acknowledgments

I want to first thank my advisor, Dr. Adam Abate, for his years of mentorship. Adam's leadership style is one that encourages intellectual creativity of the individual. With Adam, I could always be guaranteed that my ideas would be heard with an open ear and judged fairly for their scientific potential. My freedom to pursue independently-conceived projects led to some pitfalls, yes, but these are vastly outnumbered by my fruitful discoveries. I am a better scientist for paving my own way under Adam's guidance. As is often repeated by ailing sports franchises on the brink of a turnaround, sometimes you just need to "respect the process."

Freeman Lan, a former graduate student in the lab, played a big role in shaping my early graduate career. I have vivid memories of sitting by Freeman's side in Adam's office, listening to conversations I hadn't yet developed the scientific knowledge to understand. Freeman's passion for science is apparent in the way he challenged the scientific status quo. He always questioned *why* researchers did things the way they do, not simply *how*. I carry many of Freeman's philosophies and practical lab advice with me to this day.

Cyrille Delley and Harish Vasudevan, two current members of the Abate Lab, were instrumental for their work on the DAb-seq project. Yet, they are more than proficient scientists, they are genuinely great people. I am fortunate to have been able to work alongside them both.

Lastly, I want to thank my friends, family, and girlfriend, Annie. I am grateful for their love and support throughout the years. To all those who repeatedly asked, "Ben, are you done with your Ph.D.?" I can finally answer with a resounding "yes."

Contributions

Elements of this dissertation have been published elsewhere or are in preparation for publication at a peer-reviewed journal. Chapter 2: was published in *Nature Biotechnology* under the title “SiC-Seq: Single-cell genome sequencing at ultra high-throughput with microfluidic droplet barcoding[1].” A description of the single-cell barcoding technology is also featured in a peer-reviewed article published in the *Journal of Visualized Experiments* under the title “An Ultrahigh-throughput Microfluidic Platform for Single-cell Genome Sequencing[2]”; this publication was not reproduced in this thesis. Chapter 3: was published in *Methods in Cell Biology* under the title “Direct quantification of *EGFR* variant allele frequency in cell-free DNA using a microfluidic-free digital droplet PCR assay[3].” Chapter 4: has been submitted for publication in a peer-reviewed journal[4]. For brevity, some supplemental materials from these publications are not included in this thesis and are available in the online versions of these papers.

Single-cell droplet microfluidics for metagenomics and cancer multiomics

Benjamin Demaree

Abstract

Cellular heterogeneity is inherent to many biological systems, across both normal and disease states. For example, diverse ensembles of microbes in the natural environment fulfill distinct roles related to nutrient metabolism and gas fixation. In human cancers, genetic and phenotypic heterogeneity is observed among cells originating from a common oncogenic clone. Understanding biological heterogeneity, whether for metabolic engineering applications or the design of cancer therapeutics, begins at the fundamental unit of the organism: a single cell. Droplet microfluidics enables analyses of single cells at a biologically relevant scale through rapid compartmentalization and manipulation of millions of parallel reactions. In this thesis, I describe the development and application of single-cell genomics platforms leveraging droplet microfluidics to interrogate many individual genomes. These technologies enable single-cell metagenomics and multiomic analysis of single cancer cells, providing new insights into the extent of cellular heterogeneity and its implications across biology.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Genomic heterogeneity within cell populations reflects the specialized functions of its constituents. Within metagenomic systems, diverse arrays of organisms perform complementary tasks and maintain a normal community structure. In the human body, cells sharing a common inherited genome differentiate into distinct lineages, each fulfilling vital physiological roles. While genomic heterogeneity is therefore an important and even necessary aspect of biology, genetic polymorphism can also lead to dysregulation and disease, as in cancer. In this thesis, I explore different facets of cellular heterogeneity through the lens of single-cell genomics technologies using microdroplets.

Shifting scientific paradigms as well as key technological advances have combined to accelerate the recent development of single-cell sequencing technology. Early single-cell sequencing publications using flow cytometry to sort and isolate cells into microwells revealed substantial genetic mosaicism among single cancer cells[5,6], and demonstrated how tumor evolution is reflected in single-cell DNA mutations[7]. Despite their small sample sizes, on the order of tens of cells per experiment, these studies reinforced the importance of analyzing genomes at the level of individual cells and motivated the development of single-cell sequencing technologies with higher throughputs. A series of publications in 2015, at the beginning of my graduate school career, were the first to demonstrate single-cell RNA sequencing at a biologically-relevant scale, analyzing thousands of cells simultaneously. The three microdroplet-based technologies, dubbed Drop-Seq[8], inDrop[9], and Hi-SCL[10], showed that individual cells from a section of human tissue have vastly different transcriptional profiles, depending on their identity and state. The introduction of these technologies marked a new era in single-cell genomics, where the analysis of thousands of cells is now routine. Thus, the stage was set for my graduate work in the rapidly expanding field of single-cell sequencing.

The development of microfluidic technology and decrease in sequencing costs have both contributed to the growth of the single-cell genomics field. The DropSeq, inDrop, and Hi-SCL platforms, as well as the most successful commercial solution to date, produced by 10X Genomics, all use microdroplets to encapsulate and barcode cells. Microdroplets are small (~40 μm), highly monodisperse water-in-oil emulsions capable of compartmentalizing cells and other biological reagents. The rapid and serial nature of droplet generation, producing thousands of droplets per second from a single microfluidic nozzle, enables biological reactions, such as polymerase chain reaction (PCR) or reverse transcription, to be carried out in a highly-parallelized fashion. In addition to advances in microdroplet technology, the precipitous decline in sequencing cost, coinciding with the advent of next-generation sequencing (NGS) (**Figure 1.1**), accelerated the expansion of single-cell sequencing. Whereas Sanger sequencing is low-throughput and prohibitively expensive at scale, NGS can read millions of DNA fragments across thousands of cells in parallel, thereby pairing the throughput of droplet-based platforms with a complementary digital readout in the form of sequencing data.

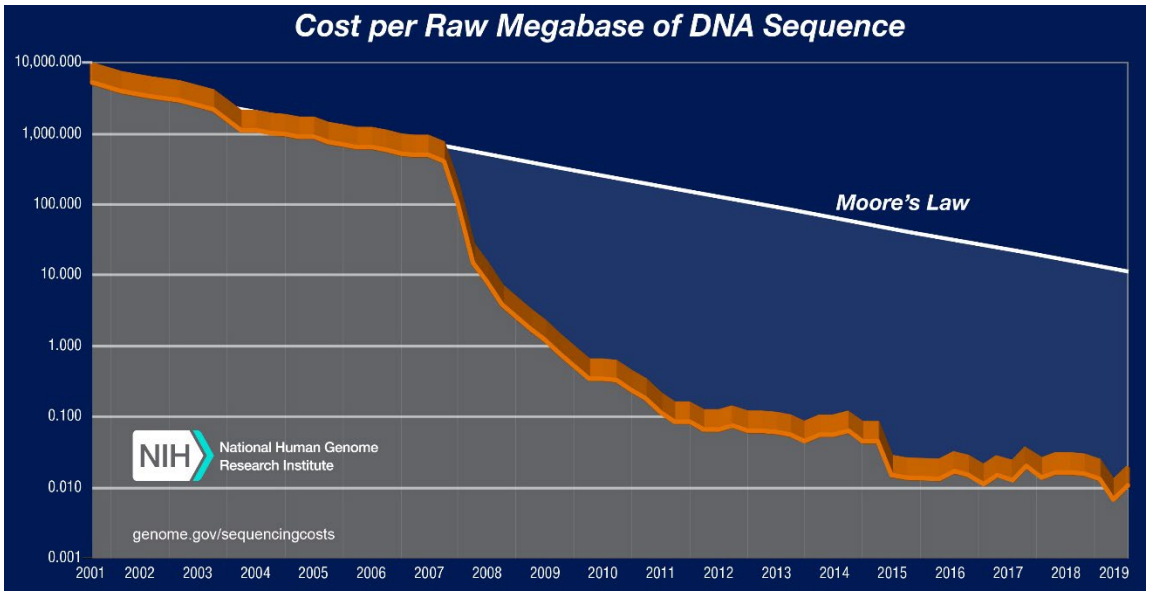

Figure 1.1: Cost per raw megabase of DNA sequencing over time.
Adapted from NIH data[11].

In the chapters that follow, I describe the development and application of single-cell genomics platforms based on droplet microfluidic technology.

Chapter 2: describes SiC-seq (**Si**ngle-**C**ell **Seq**uencing), a technology for capturing whole microbial genomes at throughputs of >10,000 cells per experiment. We demonstrate a novel microfluidic workflow wherein single cells are compartmentalized into microgels, processed, and barcoded. SiC-seq data can be used to link microbial genes to host identity, generating a network of antibiotic resistance and virulence factor distribution for a freshwater community.

Chapter 3: details particle-templated emulsification (PTE), a method for rapidly emulsifying reagents into monodisperse droplets containing a hydrogel bead. Unlike microfluidic-based approaches to emulsification, PTE requires minimal specialized equipment and its processing time is independent of initial sample volume, emulsifying billions of reactions in a matter of seconds. We apply this method to a digital droplet PCR (ddPCR) assay, demonstrating that PTE-based ddPCR can accurately quantify the variant allele frequency of a mutation in the *EGFR* gene.

Chapter 4: describes a novel single-cell multiomic technology, dubbed DAb-seq (**D**NA and **A**nti**b**ody **Seq**uencing). DAb-seq simultaneously performs targeted DNA amplification and protein counting in thousands of single cells, directly linking DNA mutations to corresponding cell immunophenotypes. Applying this technology to samples from patients with leukemia, we demonstrate that proteogenomic decoupling is prevalent in these cancers, both within and across individuals.

Chapter 2: SiC-Seq: Single-cell genome sequencing at ultra high-throughput with microfluidic droplet barcoding

2.1. Abstract

The application of single-cell genome sequencing to large cell populations has been hindered by technical challenges in isolating single cells during genome preparation. Here we present single-cell genomic sequencing (SiC-seq), which uses droplet microfluidics to isolate, fragment, and barcode the genomes of single cells, followed by Illumina sequencing of pooled DNA. We demonstrate ultra-high-throughput sequencing of >50,000 cells per run in a synthetic community of Gram-negative and Gram-positive bacteria and fungi. The sequenced genomes can be sorted *in silico* based on characteristic sequences. We use this approach to analyze the distributions of antibiotic-resistance genes, virulence factors, and phage sequences in microbial communities from an environmental sample. The ability to routinely sequence large populations of single cells will enable the de-convolution of genetic heterogeneity in diverse cell populations.

2.2. Introduction

Organisms are living expressions of their genomes and, hence, genome sequencing is a powerful way to study how they grow and function. Organisms are phenotypically diverse. This diversity is mirrored by heterogeneity at the genomic level and plays important roles in populations as a whole, particularly among populations of single cells. A common challenge when applying single-cell sequencing to heterogeneous systems is that they often contain massive numbers of cells: a centimeter-sized tumor can contain hundreds of millions of cancer cells[12], while a milliliter of seawater can contain millions of microbes[13]. Moreover, each cell has a tiny quantity of DNA, making it challenging to accurately amplify and sequence single cells. Indeed, a long history of methods based on optical tweezers[14], flow sorting[15], microfluidics[16,17], and single-cell isolation using gel matrices[18–20] have been used to isolate and process hundreds of single cells for sequencing, but this constitutes a minute fraction of most communities. The

sparseness of the sampling limits the questions that can be addressed, with the majority of findings relating to the most abundant subpopulations. A method that could markedly increase the number of cells sequenced at the single-cell level would have an impact on a broad range of problems across biology where heterogeneity is important.

Droplet microfluidics enables millions of independent picoliter reactions, and has recently been used to deep-sequence single DNA molecules[21], tag nucleosomes to enable single-cell chromatin immunoprecipitation (ChIP)-seq[22], and to profile the transcriptomes of single cells, all at high throughput[8–10]. However, sequencing the genomes of single cells presents unique challenges because genomic DNA must be purified from the cellular matter and processed through a series of enzymatic steps to prepare it for sequencing. Consequently, while droplet microfluidics provides the potential for sequencing of single-cell genomes at ultra-high-throughput, no approach for accomplishing this has yet been reported.

We describe a method for single-cell genome sequencing at ultra-high-throughput (SiC-seq) using droplet microfluidics. In SiC-seq, we encapsulate cells in hydrogel microspheres (microgels) that are permeable to molecules with hydraulic diameters smaller than the pore size, including enzymes, detergents, and small molecules, but sterically trap large molecules such as genomic DNA[23]. This allows us to use a series of washes on encapsulated cells, to perform the requisite steps of cell lysis and genome processing, while maintaining compartmentalization of each genome. Using a combination of microgel and microfluidic processing steps, we lyse the cells, fragment the genomes, and attach unique barcodes to all fragments, in a workflow that processes >50,000 cells in a few hours. The barcoded fragments for all cells can then be pooled and sequenced, and the reads grouped by barcode, providing a library of single-cell genomes that can be subjected to additional downstream processing, including demographic characterization and *in silico* cytometry (**Figure 2.1**: Schematic of SiC-seq workflow.).

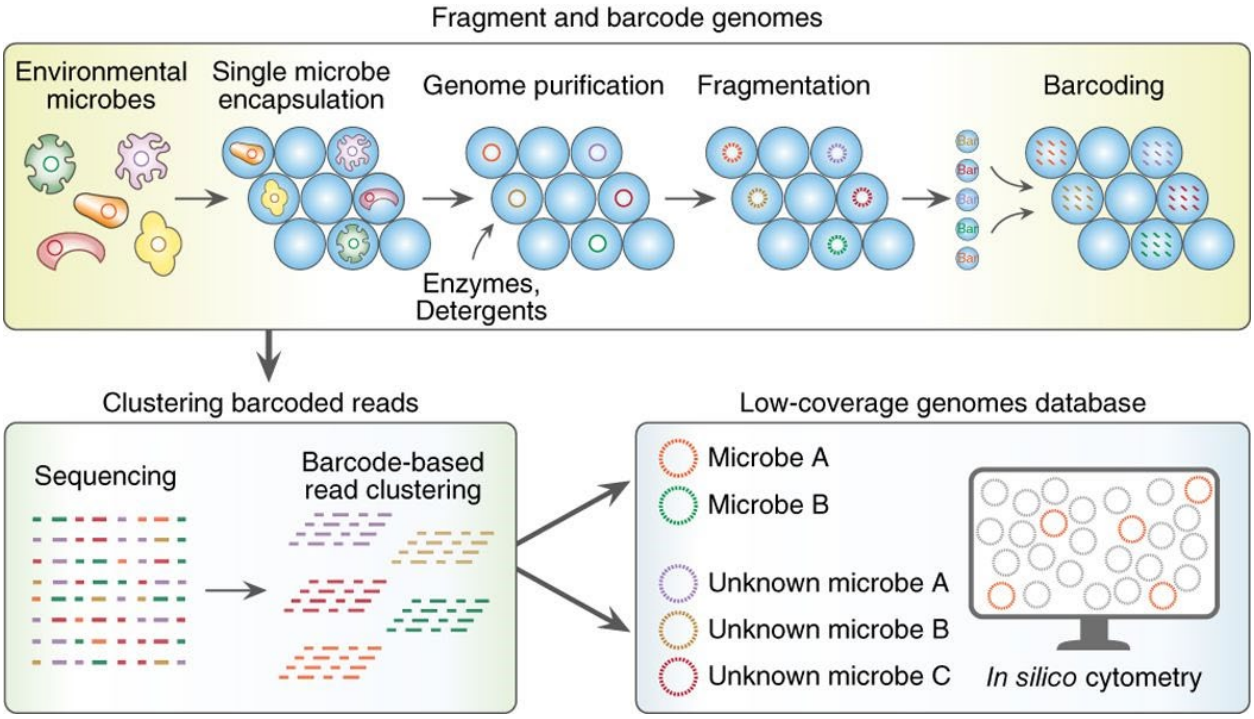

Figure 2.1: Schematic of SiC-seq workflow.
Top: droplet workflow to generate single-cell genome-barcoded sequencing library. Bottom left: sequencing and generation of barcode groups representing reads from single cells. Bottom right: the groups of reads comprise a database of low-coverage genomes of single cells, which can be searched repeatedly *in silico*.

2.3. Results

2.3.1. SiC-seq workflow

The principal strategy of SiC-seq is to label all DNA fragments originating from the same genome with a sequence identifier (barcode) unique to that cell. The resultant products are chimeric, comprising a barcode sequence covalently linked to a random fragment of the cell genome. The barcodes allow all reads belonging to a given cell to be identified through shared sequence. We use libraries of barcode droplets containing the barcode sequences, which we merge with the genome-containing droplets to be barcoded[21]. To prepare a barcode droplet library, we encapsulated into droplets, at limiting dilution, oligonucleotides comprising 15 random bases flanked by constant sequences with PCR reagents and primers complementary to the constant regions of the barcodes with one side containing the Illumina P7 flow cell

adaptor[24] (**Figure 2.2a**). The droplets were then thermal-cycled to amplify the barcode sequences via digital droplet PCR, generating ~10 million barcode droplets in a few hours.

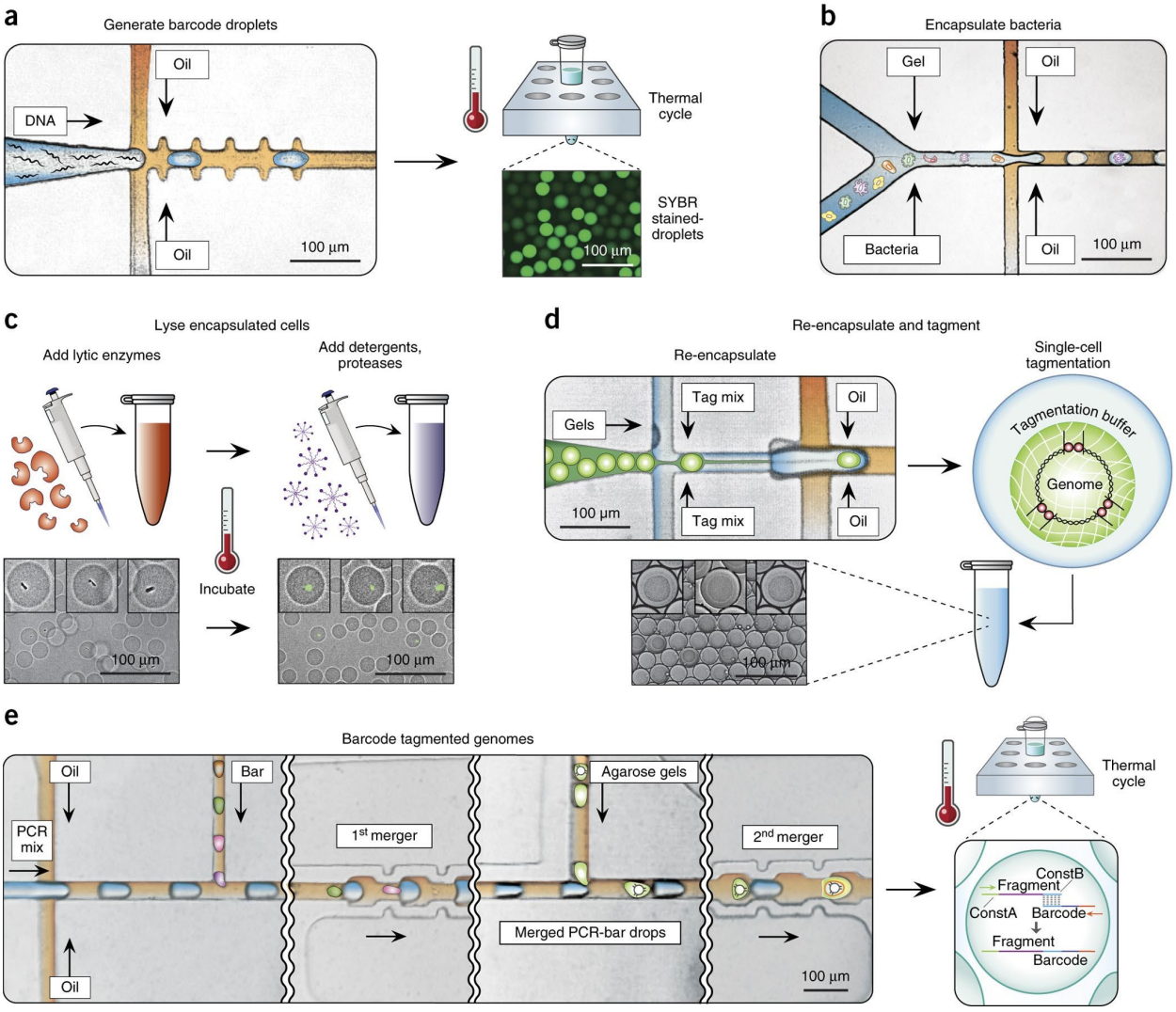

Figure 2.2: Microfluidic and biochemical workflow to generate a SiC-seq library.
(a) Generating barcode droplets by encapsulating random DNA oligos at limiting dilution and amplification by in-droplet PCR (SYBR-stained for visualization). **(b)** Cells are encapsulated at limiting dilution with molten agarose to generate agarose microgels each containing a single cell. **(c)** The single-cell genomes are purified through a series of bulk enzymatic and detergent lysis steps. **(d)** Microgels are re-encapsulated in droplets containing tagmentation reagents. **(e)** The droplets containing tagmented genomes are merged sequentially with PCR reagents and barcode droplets at a 1:1 ratio, followed by PCR to splice barcodes to genomic fragments.

Before the single-cell genomes can be barcoded, they must be physically isolated, purified, and fragmented. To accomplish this, we encapsulated single cells in agarose microgels using a two-stream co-flow droplet maker, which merges a cell suspension stream with a molten agarose stream, forming a droplet consisting of an equal volume of both streams (**Figure 2.2b** and **Figure 2.3a**). The droplet maker runs at ~10 kHz, allowing us to generate ~10 million ~22-μm-diameter droplets in ~20 min in a total volume of aqueous emulsion of ~60 μL. Hence, droplet generation is fast and the total volume consumed small, allowing us to load cells at a rate of 1:10 to minimize multi-cell encapsulation. After solidifying the agarose by cooling, the microgels were then transferred from oil to aqueous carrier phase to be subjected to cell lysis and genome purification. To lyse the cells, we incubated the microgels overnight in a mixture of lytic enzymes, digesting the protective microbial cell walls. We then incubated them in a mixture of detergents and proteases for 30 min, solubilizing lipids and digesting proteins, preserving only high-molecular-weight genomic DNA, which we verified by staining with SYBR green dye (**Figure 2.2c**). To fragment the genomes and attach the universal sequences to act as PCR handles, we re-encapsulated the gels in the Nextera reaction (**Figure 2.2d** and **Figure 2.3b**). Because the transposases are dimeric, the fragmented genome remained intact as a macromolecular complex, remaining sterically encased within the hydrogel network[25] (**Figure 2.4**). Nevertheless, we re-encapsulated the gels into separate droplets during fragmentation to ensure that there was no cross-contamination of DNA between the gels.

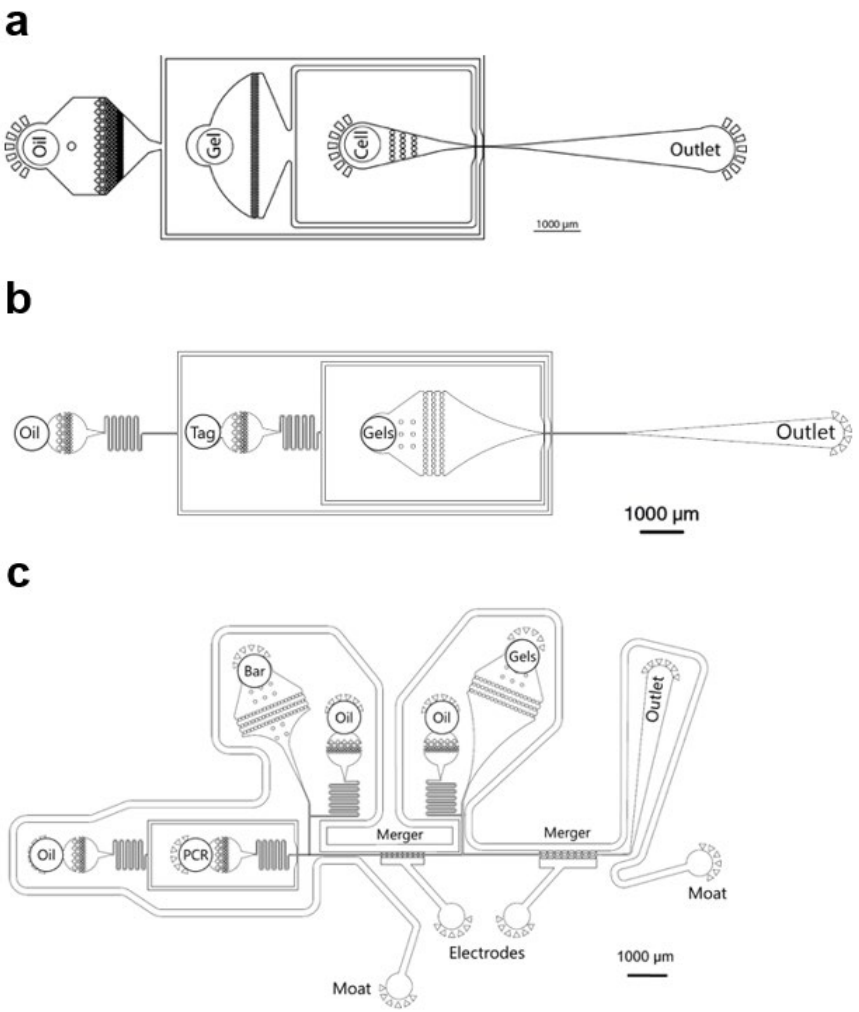

Figure 2.3: Drawings of SiC-seq microfluidic devices.
Schematics of microfluidic devices used to: **(a)** generate barcode droplets and encapsulate cells in agarose microgels; **(b)** re-encapsulate gels in tagmentation reagents; **(c)** merge gel droplets with barcode droplets and PCR droplets.

After the genomes were purified and fragmented, they were barcoded for sequencing. We used a microfluidic device that merged each microgel-containing droplet with droplets containing PCR reagents and a barcode droplet (**Figure 2.2e** and **Figure 2.3c**). The resulting droplets, which contained fragmented-genome and barcoded DNA, were collected into a PCR tube and thermal-cycled, splicing the barcode sequences onto the genomic fragments via complementarity through the PCR handles added by the transposase. At this point, the spliced fragments contained both the P5 and P7 Illumina sequencing adaptor required for sequencing

on the Illumina platforms. We removed droplets that coalesced during thermal cycling using a micropipette; the remaining droplets were chemically merged and their contents pooled and prepared for sequencing. After sequencing, the reads were filtered by quality and grouped by barcode, providing single-cell genomic sequence data.

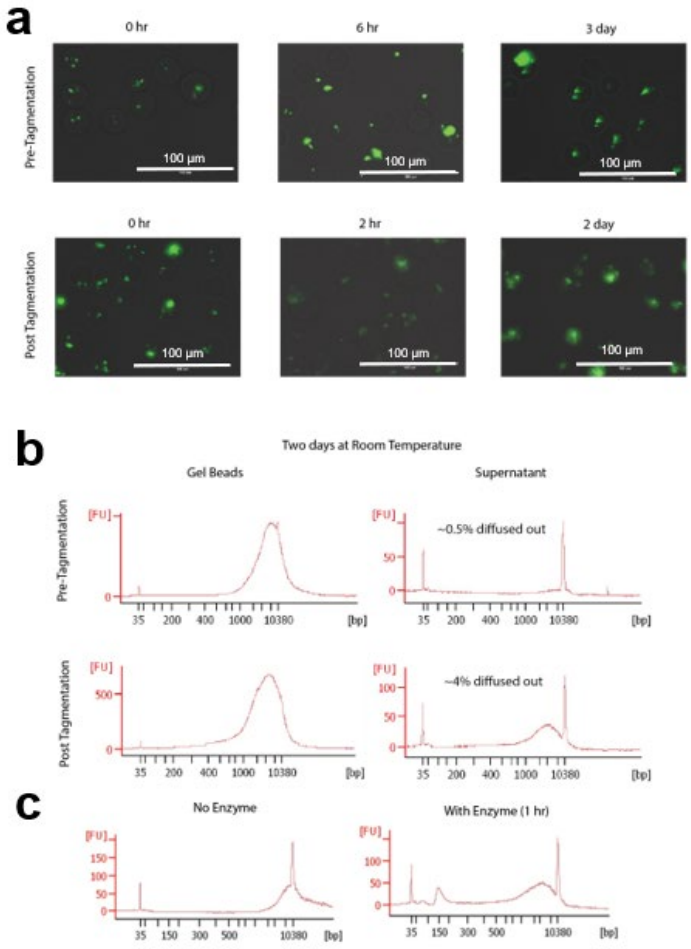

Figure 2.4: Characterizing diffusion of genomic fragments inside agarose microgels.
(a) SYBR staining was used to monitor diffusion of genomes in microgels before and after tagmentation. **(b)** After two days at room temperature, the beads were pelleted by centrifugation and DNA was extracted from the beads and the supernatant and quantified using the Qubit dsDNA high sensitivity assay and bioanalyzer high sensitivity chip. The shift in fragment size is relatively minor as a result of the relatively low stoichiometric ratio of transposase to genome used. **(c)** Encapsulated genomes are reacted with a higher stoichiometric ratio of transposase to genome are visualized on a bioanalyzer high sensitivity chip to show fragmentation efficiency of the gel encapsulated genomes.

2.3.2. *Validation of SiC-seq on an artificial microbial community*

The objective of SiC-seq is to provide single-cell genomic sequences bundled in barcode groups. To validate that SiC-seq generated single-cell barcode groups, we applied it to an artificial microbial community containing three Gram-negative bacteria, five Gram-positive bacteria, and two yeasts, which are typically difficult cell types to lyse. We prepared a single-cell library from this community using SiC-seq and sequenced it on an Illumina MiSeq, yielding ~6 million single-end reads of 150 bp after quality filtering. We grouped reads by barcode and discarded groups with <50 reads, yielding the final 48,989 barcode groups (**Figure 2.5a**). Each barcode group represented a low-coverage genome of a cell, with a sequencing depth of ~0.1% to ~1% (**Figure 2.6**).

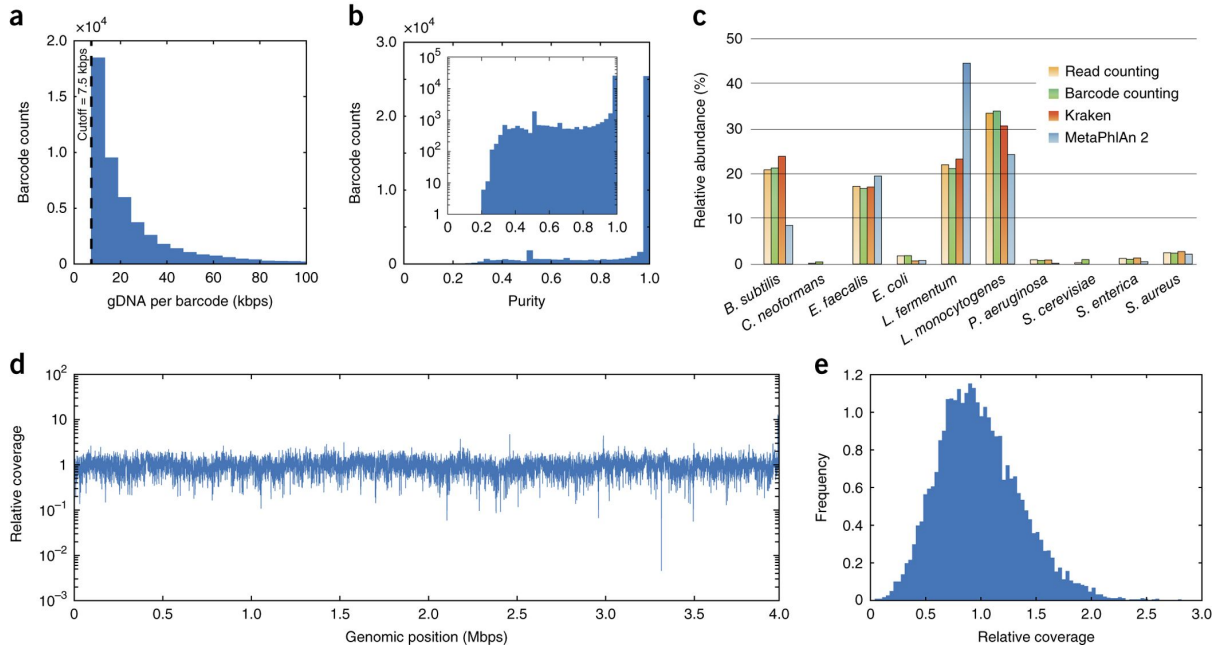

Figure 2.5: SiC-seq performance on an artificial microbial community consisting of ten different cell species.
(a) Distribution of sequencing yield of each barcode group. **(b)** Histogram of the purity of each barcode group, which is defined as the fraction of reads mapping to the most mapped species for that group. The inset is plotted with the counts on a logarithmic scale. **(c)** Relative abundance estimates of each species using read counting, barcode counting, and two different taxonomic profiling programs (Kraken and MetaPhlAn 2). **(d)** Relative coverage of the *Bacillus subtilis* genome for all *B. subtilis* barcode groups, showing good uniformity. **(e)** Coverage histogram for the B. subtilis genome binned by relative coverage.

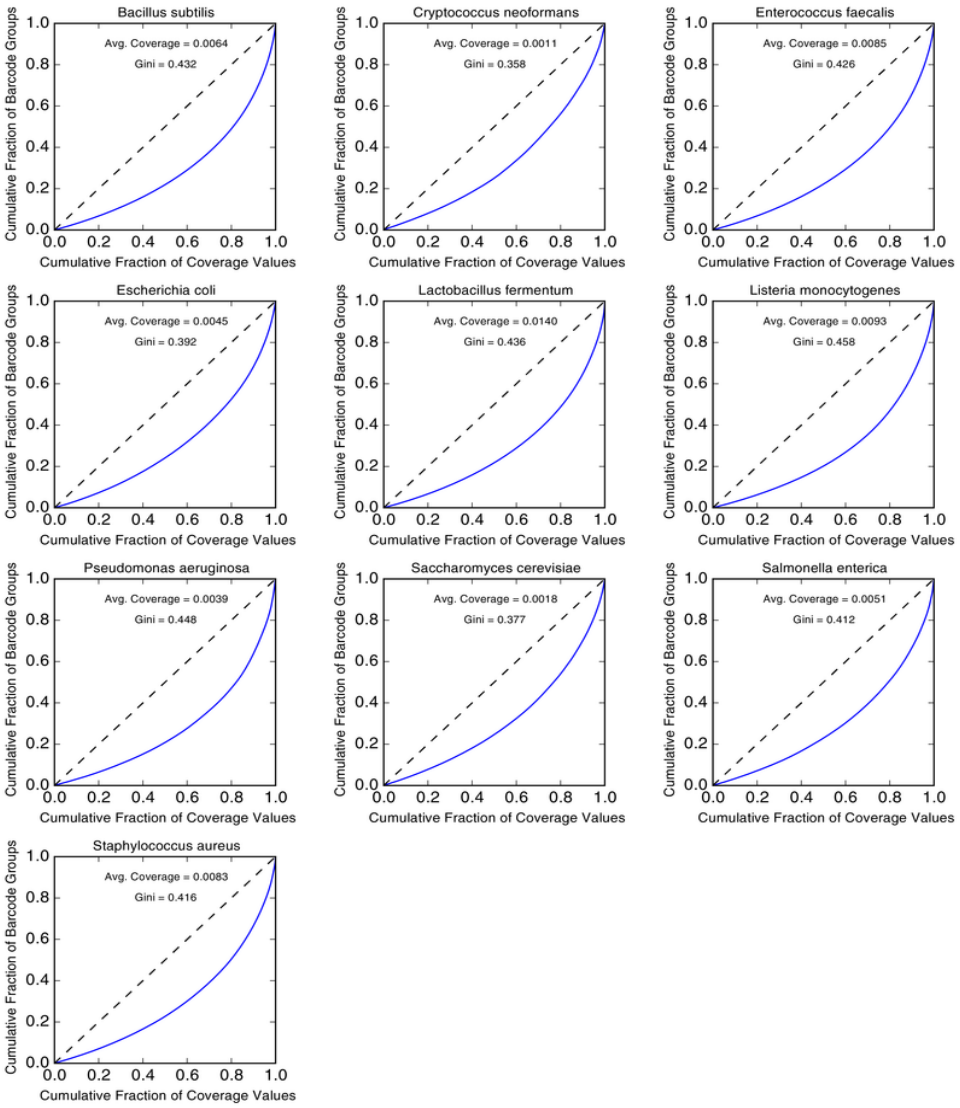

Figure 2.6: Lorenz curves of barcode group coverage, by species.
The average depth and distribution of genome coverage of each barcode group plotted as a Lorenz curve for each species in the 10-cell control experiment.

To determine whether the barcode groups indeed corresponded to single cells, we mapped all reads to the reference genomes of the ten species. If two microbes reside within the same barcode group, reads will map to two genomes. We defined a group purity score as the fraction of reads mapping to the most mapped reference (the ideal barcode group has a purity score of 1.0). The distribution of purity scores was strongly skewed to high values with the majority of purity scores >0.95, suggesting that most barcode groups represented single cells;

this result was consistent even when we took into account the different genome sizes of the ten species (**Figure 2.5b** and **Figure 2.7**) and when we examined purity individually for each species (**Figure 2.8**). We further examined the rare barcode groups with low (<0.8) purity scores and determined that the majority of those barcode groups represent rare cases where two cells were encapsulated into one droplet or the occasional coalescence of two single-cell-containing droplets (**Figure 2.9**).

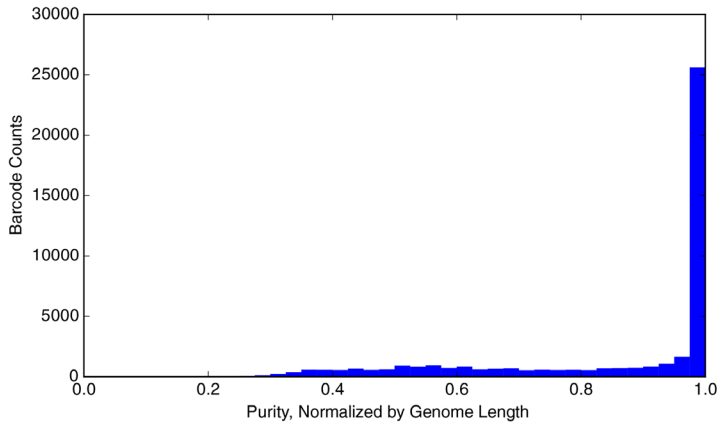

Figure 2.7: Genome size-normalized barcode group purity scores.
Genome size-normalized purity scores of barcode groups in the 10-cell control experiment. Genome size-normalized purity scores are calculated using the same method using the fraction of the genome sequenced for each respective species rather than the raw number of reads.

To determine whether the abundance of SiC-seq barcodes reflected the abundance of corresponding organisms in the data set, we compared abundance estimates calculated by short-read alignment, taxonomic profiling programs, and counting under brightfield microscopy (**Figure 2.5c** and **Figure 2.10**). We found that all methods were in reasonable agreement when reads were pooled and analyzed in bulk and when species identities were assigned to each barcode based on the most commonly mapped species in a group. This demonstrates that SiC-seq enables estimation of species abundance in a microbial population consistent with accepted metagenomic methods.

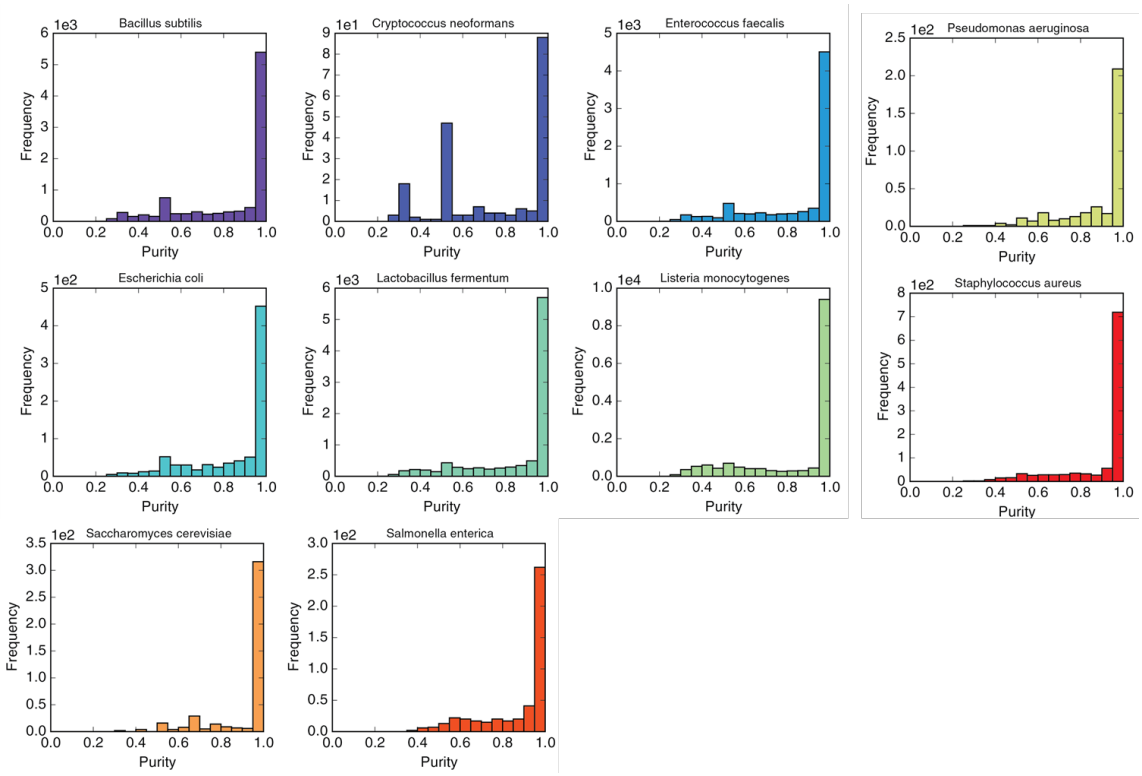

Figure 2.8: Barcode group purities, by species.
Purity scores of barcode groups separately plotted for each species in the 10-cell control experiment.

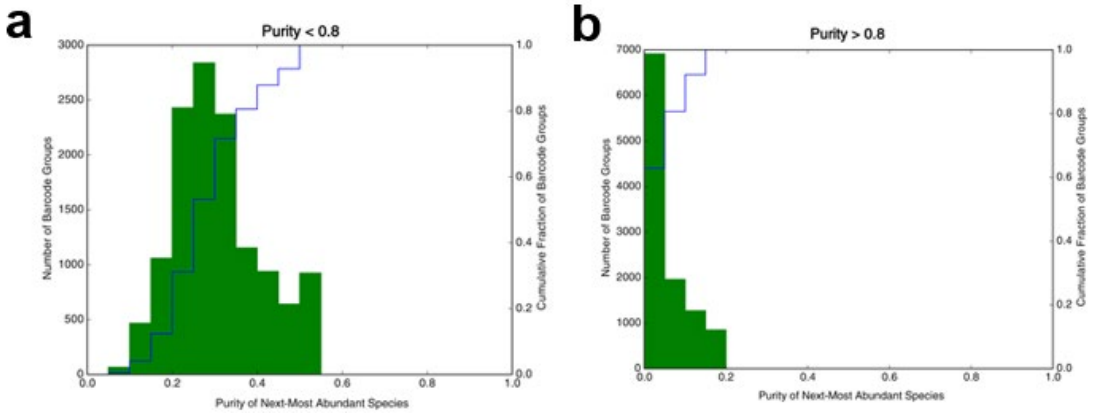

Figure 2.9: Barcode group purity scores for second-most abundant species.
Purity scores of the next-most abundant species in a) barcode groups of purity <80%; b) barcode groups of purity >80%. In barcode groups with <80% purity, the purity scores of the next-most abundant species tend to be high from ~20% to 50%, reflecting that those two species represent the majority of the reads in the barcode group, suggesting that these barcode groups represent double encapsulations. Barcode groups with 100% purity are not represented in the plots. Blue line represents cumulative barcode counts normalized to 1.

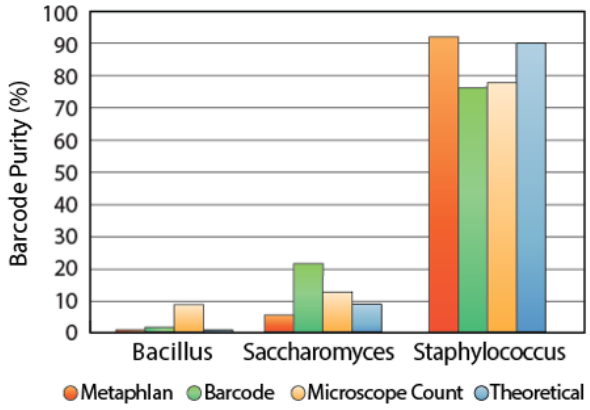

Figure 2.10: SiC-seq performance on an artificial 3-cell microbial community.
Relative abundance estimates of each species are calculated using barcode counting (Barcode), marker gene counting without barcodes (Metaphlan), and manual counting under the microscope after cell encapsulation (Microscope count) and while in culture (Theoretical).

Sequencing the genome of a single cell typically incurs coverage distribution bias[26] due to uneven amplification of DNA starting from a single genome copy. To investigate coverage distribution bias in SiC-seq, we plotted the normalized coverage distribution for reads aggregated from all barcode groups for each microbe (**Figure 2.5d,e** and **Figure 2.11**). With the exception of coverage gaps due to the low abundance of cells of certain species within the standard microbial community, we observed no substantial coverage bias. This indicates that the sampling of each genome within a barcode group was random, so that when all groups were overlaid, a uniform distribution was obtained. We further inspected the distribution of reads in individual barcode groups and found no substantial bias. We believe that coverage bias was minimal because each genome was amplified in a tiny volume of ~65 pL, which has been shown to curtail bias-inducing runaway of exponential amplification[27]. Since the sequencing library was composed of ~50,000 amplified genomes, the amplification of each genome was limited by the tiny volume while still producing sufficient total DNA for sequencing.

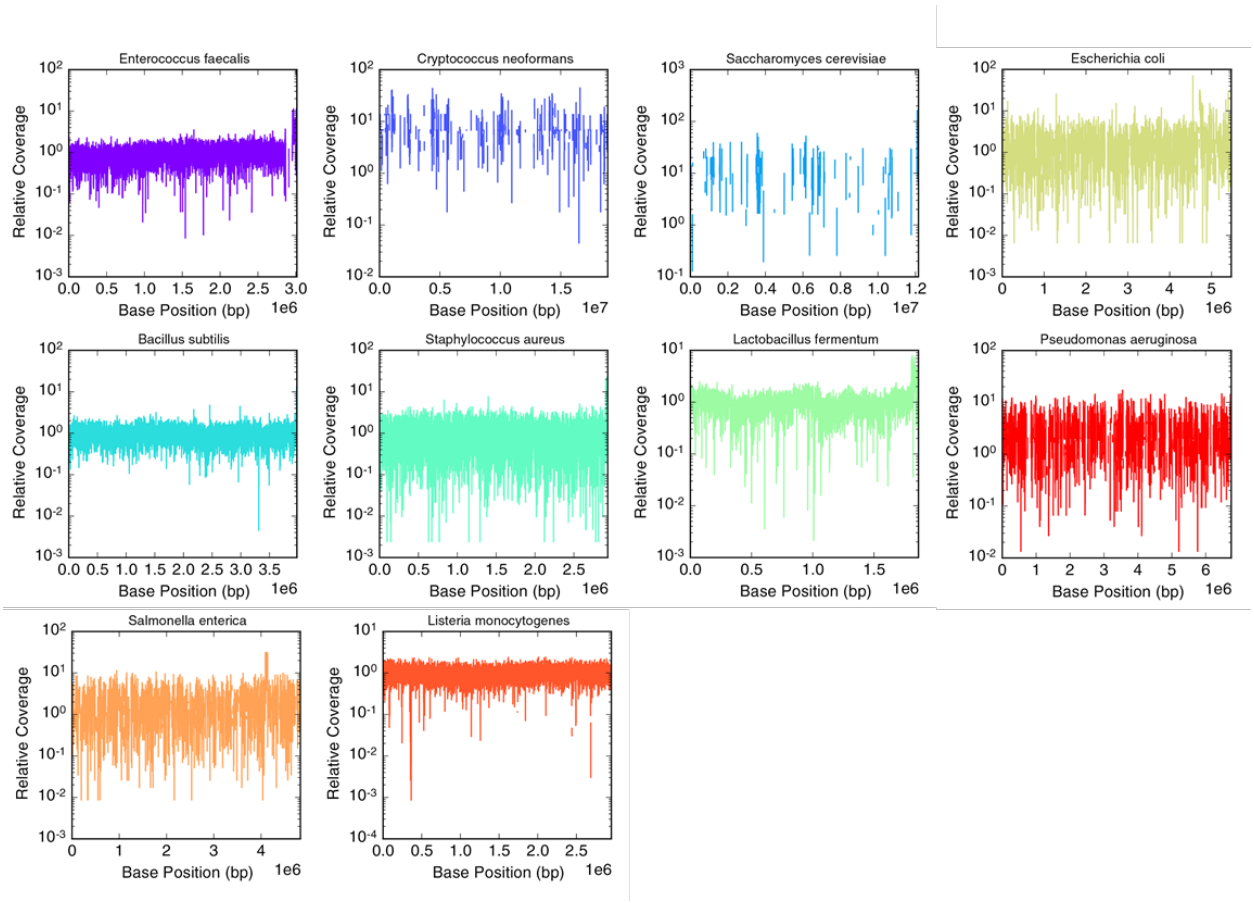

Figure 2.11: Aggregate genomic coverage of all the barcode groups for species in the synthetic microbial community.
Species at low abundance show frequent dropouts characterized by dips in the graph, but instances of systematic bias characterized by sharp peaks are rarely observed.

2.3.3. SiC-seq data analysis with in silico *cytometry*

The genomic sequences generated using SiC-seq are grouped by cell barcodes, which is complementary to the sequences generated from shotgun metagenomic sequencing. Existing computational tools are ill-suited to analyze these data because they do not exploit the single-cell barcode information unique to SiC-seq. To address this, we utilized a sequence analysis pipeline in which reads are organized hierarchically as barcode groups, generating a single-cell-reads database (SiC-Reads). To build SiC-Reads, we filtered raw sequences by quality, grouped them by barcode, and estimated a taxonomic classification of each group using phylogenetic profilers. We also estimated a purity score equal to the fraction of reads mapping

to the dominant taxon within the classifiable set. Additional properties of barcode groups and reads, such as presence of sequences corresponding to antibiotic-resistance genes, can be added to the database as they are discovered during analysis.

The massive set of single-cell genomes present in SiC-Reads provides new opportunities for discovering associations between sequences within single cells, in a process we dub *in silico* cytometry. SiC-Reads comprises a collection of single-cell genomes that can be sorted *in silico*, analogous to what is commonly done with flow cytometry on single cells. The database can be sorted repeatedly to mine for correlations between different genetic sequences and structures. Moreover, as new associations are learned, new sorting parameters can be defined, enabling discoveries without having to repeat the experiment.

2.3.4. Taxonomic distribution of antibiotic resistance in microbes

To demonstrate *in silico* cytometry, we used SiC-seq to sequence a microbial community recovered from coastal seawater of San Francisco. We obtained ~8 million reads of 150-bp length after quality filtering (representing ~55% of raw reads), with which we generated a SiC-Reads database. Using a phylogenetic profiler, 601,348 (6.89%) of reads were successfully classified into taxa representing 99.8% bacteria, 0.04% archaea, and 0.16% viruses (**Figure 2.12a**). Barcode groups were assigned a taxonomic classification based on the reads they contained, following the rule that more than 10% of reads in a barcode group must be classified, and the assigned classification is the taxon with the most supporting reads. Most barcode groups were high purity based on the classifiable sequences (~91% average), in accordance with our control sample (~94% average) (**Figure 2.12b**). Using this SiC-Reads database, we demonstrated *in silico* cytometry by exploring the distribution of antibiotic resistance, virulence factors, and phage sequences in the microbial community.

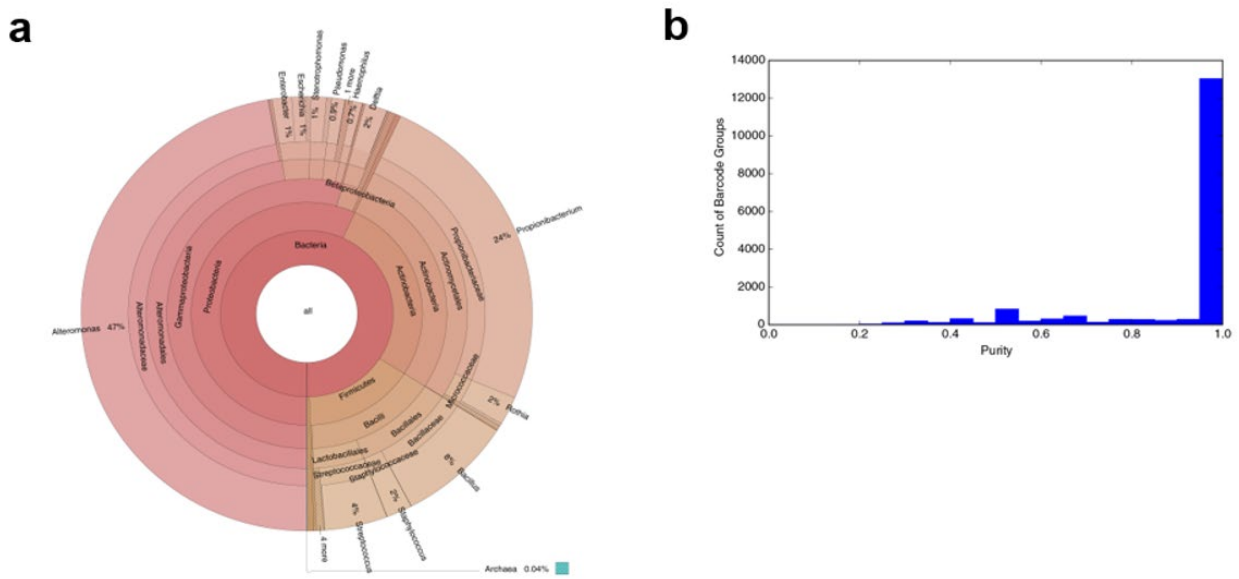

Figure 2.12: Analysis of the marine microbial community used to demonstrate *in silico* cytometry.
(a) Taxonomic abundance of the SiC-Reads database by barcode groups; **(b)** Distribution of purity of barcode groups in the database at the genus level.

Antibiotic resistance has become increasingly common and represents a considerable threat to global human health[28]. While antibiotic-resistance genes can be identified in most environments by short-read sequencing, scant information on how they are distributed among taxa is available, because obtaining this information usually requires testing or whole genome sequencing of single species; however, culture conditions for most species have not been identified, precluding such analyses.

To determine the distribution of antibiotic-resistance genes among taxa in our data set, we searched our SiC-Reads database for known antibiotic-resistance genes, finding 1,081 (0.012% of reads), representing 108 (0.30%) barcode groups. The taxonomic distribution of antibiotic-resistance genes in our database had a clear structure, although it did not correlate with what is known from genomes in public databases (**Figure 2.14a** and **Figure 2.13a**). This is unsurprising as differences are expected in the natural coastline environment compared to the environment of isolated and sequenced strains. The most abundant taxa associated with

antibiotic resistance were not the most abundant community members overall, suggesting that in this community, certain taxa tend to associate more with antibiotic-resistance genes.

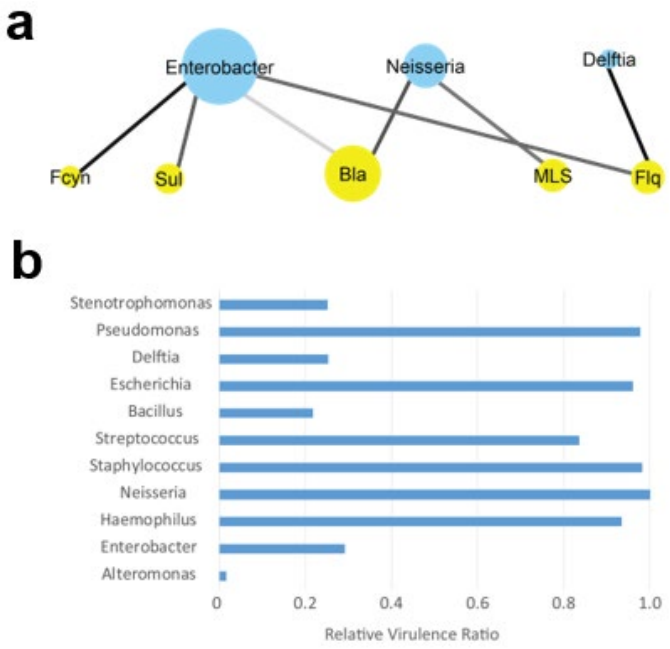

Figure 2.13: Simulation of sequencing data from marine strains.
Reference data obtained by simulating reads from genomic sequences of isolated strains for comparison against data in the marine microbial community. **(a)** Antibiotic resistance network for whole genome sequenced strains in public databases; **(b)** Virulence factor ratios calculated for publically available strains.

2.3.5. Association of virulence factors with host bacteria

Virulence factors, like antibiotic-resistance genes, are important genetic factors in determining the threat that specific microbes pose to human health. Many opportunistic pathogens reside in natural communities in the environment and cause outbreaks when transmitted to a suitable host[29]. Monitoring and detecting potentially pathogenic microbes is important for public health. Like antibiotic-resistance genes, traditional methods can detect the presence of these genes but not their taxonomic distribution.

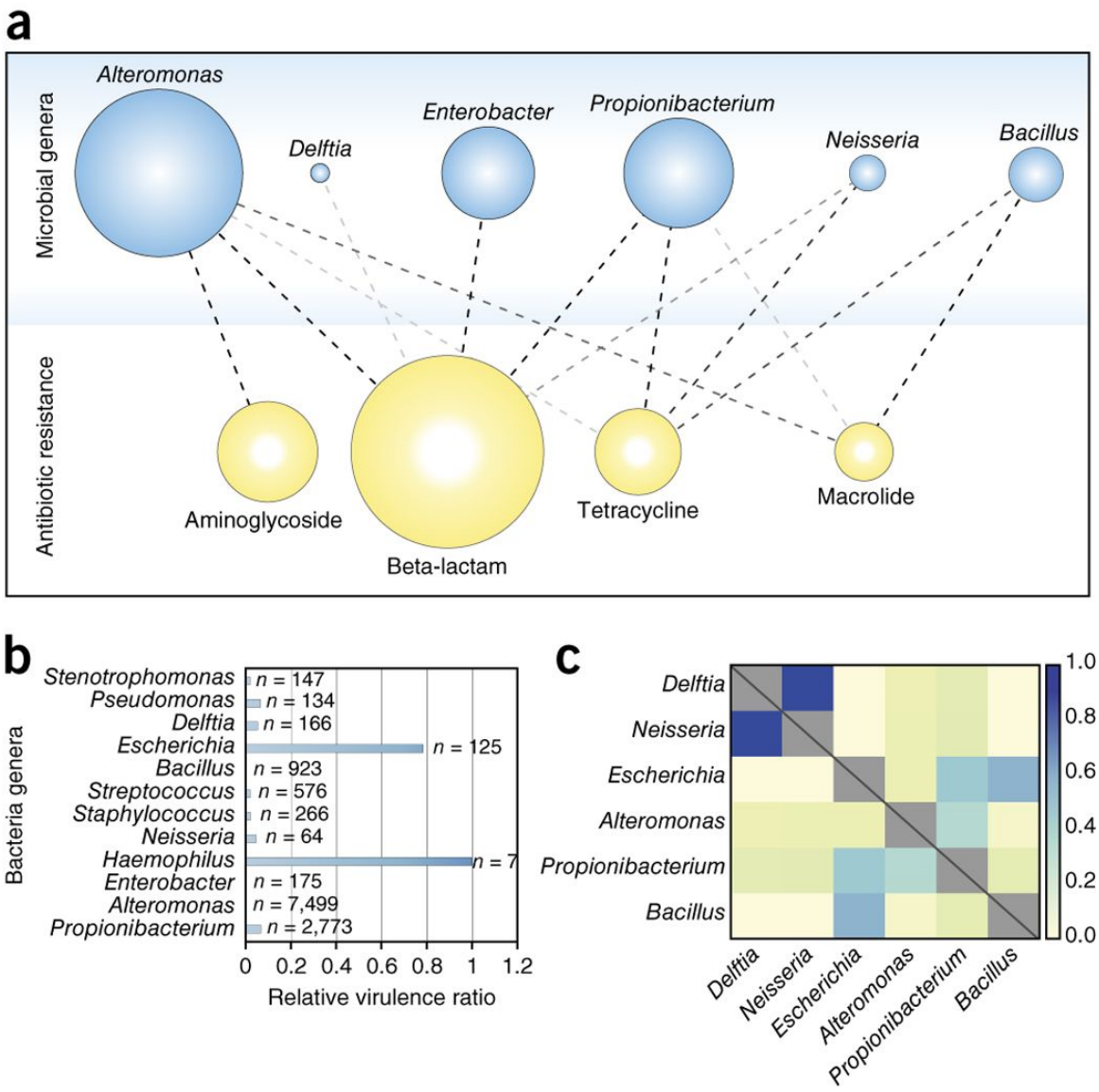

Figure 2.14: Application of SiC-seq to a marine community recovered from the San Francisco coastline.
(a) Distribution of antibiotic-resistance genes according to genus of host microbe. The more opaque the lines connecting the circles, the greater the number of interactions detected in the database. **(b)** Relative abundance of virulence factors in each genus detected in the community. **(c)** Relative potential for transduction between bacterial taxa, determined by the relative number of common phage sequences detected in their respective genomes, plotted as a heat map.

To examine the taxonomic distribution of virulence factors in our data set, we searched our coastal microbial community database for known virulence factor genes, yielding matches in 1,949 (0.022%) reads in 101 (0.28%) barcode groups, consisting of 29 prevalent virulence

factors distributed among 13 microbial genera. The abundance of each taxa where virulence factors were found did not reflect their abundance in the total population, suggesting that certain genera tend to carry more virulence factors than others. To quantify this, we calculated the virulence factor ratio, the ratio between the number of barcode groups containing virulence factors and the total number of barcodes in the community for that species, and normalized the results to the highest virulence factor ratio for comparison (**Figure 2.14b**). *Haemophilus* and *Escherichia* stand out among all species; both are known opportunistic human pathogens.

Comparing the virulence factor ratios of the San Francisco coastline community with ones calculated for publicly available whole genomes and downsampled to match our per-cell read depth (**Figure 2.13b**), we found that the ratios were higher for the public genomes, an expected result given that isolated and sequenced genomes are biased toward pathogenic strains.

2.3.6. Determining transduction potential between bacteria

Many virulent bacterial strains are thought to arise from horizontal gene transfer aided by cross-infection of bacteriophages. Phages can modify the genomes of their hosts, leaving a copy of their own genome behind or transporting fragments of one species to another in a process known as transduction[30,31]. Characterizing the distribution of these mobile elements is challenging in an ecological context because confident identification of foreign genomic fragments within a specific host requires sequencing large numbers of cultures of single species or single cells. Nevertheless, this information is valuable for understanding how bacteria transfer genetic material in general, and how virulent new strains may emerge through this mechanism.

To explore transduction in the microbial community, we searched the SiC-Reads database of the San Francisco coastal community for barcode groups containing phage sequences. A phage sequence found in a bacterial genome is evidence of infection, an association that is normally extremely difficult to make for uncultivable microbes and their likely

uncultivable infecting phages. We found matches in 6,805 (0.078%) reads representing 260 (0.72%) barcode groups and 106 phage genomes. Since transduction can occur between two host cells that can be infected by the same phage, the potential for transduction depends on the likelihood of phages infecting both hosts. To visualize this, we plotted the normalized sum of the number of times we detected the sequences matching to the same phage in two bacterial taxa, normalized by the number of barcode groups in those taxa (**Figure 2.14c**). According to this analysis, *Delftia* and *Neisseria*, which are the most closely related out of the taxa in our analysis, had the highest potential for transduction. The dearth of representative phage genomes in databases and the limited sequence information per barcode group limited the accuracy of this approach. Therefore, higher coverage of the genomes and better phage genome databases are required to definitively identify the phages that are found in the database. Nevertheless, SiC-seq's ability to detect these sequences and correlate them within single genomes can provide a useful approach to study phage–host interactions.

2.4. Discussion

SiC-seq generates a metagenomic database grouped by single-cell genomes amenable to repeated mining by *in silico* cytometry, for rapid hypothesis generation and testing. We demonstrated its use in measuring the distributions of antibiotic-resistance genes, virulence factors, and transduction potential in microbial communities. The ability of SiC-seq to sequence all cells in a sample without the need to culture them should help us to characterize the microbial dark matter.

The barcoded nature of SiC-seq data necessitates additional quality control measures for the data, in addition to the quality control measures used in standard sequencing. First, the barcode reads themselves must be of high quality, thus any reads containing low-quality barcode sequences are eliminated, regardless of the quality of the genomic sequences. Second, barcode groups must be quality-controlled to remove small barcode groups, which are

the result of mutations in the barcode sequences and background contamination of free DNA. These quality control measures together result in a typical yield of ~55% of raw reads contributing to the SiC-Reads database. Improvements in yield can be made by, for example, computationally identifying reads with mutated barcodes and 'correcting' their sequence, but we have found only modest improvements in yields using this method alone[21].

The taxonomic classification of microbes remains an integral part of studying community dynamics, from ecosystems on Earth to those residing in and on our bodies[32,33]. However, the taxonomic classification of short reads is error prone, due to the diversity of microbes in most communities and the high degree of horizontal gene transfer that mixes genomic elements in unpredictable ways. SiC-seq improves upon traditional metagenomics sequencing in addressing this challenge because taxonomic identification can be made on the basis of hundreds of reads within a barcode group. Advanced strategies can be applied to infer the taxonomy of a barcode group, including Bayesian probabilistic ones based on classification of each read in the group, or ones weighted toward specific taxonomic markers. Even with this improvement, accurate classification is difficult because the vast majority of sequences remain unclassifiable and the classification of sequences are biased toward well-sequenced taxa in the databases. As genome coverage improves in future iterations of SiC-seq, taxonomic classification of barcode groups should become more confident and precise, potentially arriving at classifications down to the strain under certain circumstances. It is worth noting that taxonomic classification with SiC-seq is also subject to the fundamental limitations of reference-based classification paradigms, where the classification is only as accurate as the match between the sample and the references. Hence, like traditional methods, SiC-seq phylogenetic profiling will become more reliable and complete with the expanding database of reference genomes.

The degree of genome coverage affects the usefulness of single-cell data, including the ability to generate assemblies or identify characteristic sequences for *in silico* cytometry. A limitation of SiC-seq is that, while the number of cells sequenced far exceeds currently

described methods, the coverage per cell is substantially lower. Therefore, dropouts in coverage and false negatives can be expected in *in silico* cytometry analysis. For abundant organisms with a random distribution of coverage in each barcode group, the system is robust to dropouts because results are averaged over many barcode groups. For example, approximately 7,000 *Alteromonas* barcode groups were taken into account to determine the antibiotic-resistance profile for *Alteromonas* bacteria. However, for less abundant species, such as *Haemophilus*, more dropouts can be expected because there may not be enough total sequence information to detect a specific genetic factor. For this reason, the analysis of SiC-Read databases should be limited to relative comparisons of species within the database, and the abundance of target genes within subpopulations should be normalized to the number of barcode groups in the subpopulation. It is worth noting that the dropout phenomenon is not unique to SiC-seq data, but applies to all metagenomic sequencing data where the subpopulation to be analyzed represents a very small fraction of the whole.

Although coverage can be increased by sequencing more reads, the coverage per cell per barcode group will be <100%. This is because the method begins with a single genome copy without amplification and losses incurred during enzymatic and microfluidic processing are irrevocable, thus limiting the maximum coverage attainable. In future iterations of SiC-seq, coverage may be increased by amplifying genomes before processing, for example, with multiple displacement amplification in droplets[18]. Additionally, different strategies for barcoding genomes may yield higher coverage, such as recently described combinatorial indexing via transposase libraries[25], which should be applicable to single-cell genomes encapsulated in microgels.

The *de novo* assembly of whole genomes from metagenomics sequences is a common goal in the field of metagenomics. Mate-paired sequencing can be used to bridge contigs in a metagenomics sequencing data set and potentially assemble whole genomes given sufficient coverage[34]. Though powerful, the method is limited by the required micrograms of starting DNA,

which can be difficult to obtain from microbial ecosystems. Furthermore, many mate-paired reads are required to assemble a whole genome, since each mate-pair bridges only two contigs. SiC-seq data improve on mate-paired sequencing in this respect by requiring minimal amounts of sample as well as enabling the bridging of multiple contigs per barcode group. Consequently, SiC-seq should allow generation of draft genomes from shotgun metagenomic data with far less DNA input and sequencing effort.

While we focused on microbial communities, SiC-seq is also applicable to populations of mammalian cells, where it can have a more direct impact on human health. The grouped reads provided by SiC-seq should afford the information required to determine copy-number variations within the genome, which is relevant to cancer[35]. The enormous size of mammalian genomes, however, limits the number of cells that can be sequenced for a target level of coverage. Nevertheless, as the cost of sequencing continues to decrease, more cells can be sequenced to greater depth, creating opportunities for characterizing mammalian tissues, cell by cell.

The SiC-seq method is a means to isolate and barcode large DNA molecules, irrespective of the entity from which they originate. While we have focused on cells, similar approaches can be applied to any entities whose genomes can be trapped and processed within the gel matrix. SiC-seq's ability to build and mine large databases of genomes grouped by single cells should contribute to the characterization of heterogeneity across biology.

2.5. Methods

2.5.1. Microfluidic devices

The microfluidic devices are fabricated by pouring poly(dimethylsiloxane) (Dow Corning, Sylgard 184) over a negative photoresist (MicroChem, catalog no. SU-8 3025) patterned on a silicon wafer (University Wafer) using UV photolithography. The PDMS devices are cured in an oven for 1 h, extracted with a metal scalpel, and punched with a 0.75-mm biopsy core (World Precision Instruments, catalog no. 504529) to create inlets and outlets. Devices are bonded to a

glass slide using an oxygen plasma cleaner (Harrick Plasma), and the channels treated with Aquapel (PPG Industries) and baked at 80 °C for 10 min to render them hydrophobic.

2.5.2. Barcode emulsions

Barcode emulsions are prepared through a digital PCR process wherein barcode oligonucleotides are amplified as single molecules in droplets containing PCR reagents. Barcode oligonucleotides (GCAGCTGGCGTAATAGCGAGTACAATCTGCTCTGATGCCGCATAG NNNNNNNNNNNNNNNTAAGCCAGCCCCGACACT) (IDT) at 0.01 pM concentration are added to a PCR reaction mix containing 1× NEB Phusion Hot Start Flex Master Mix (NEB, catalog no. M0536L), 2% (w/v) Tween 20 (Sigma-Aldrich, catalog no. P9416), 5% (w/v) PEG-6000 (Santa Cruz Biotechnology, catalog no. sc-302016), and 400 nM primers FL128 (CTGTCTCTTATACACATCTCCGAGCCCACGAGACGTGTCGGGGCTGGCTTA) and FL129 (CAAGCAGAAGACGGCATACGAGATCAGCTGGCGTAATAGCG, contains P7 adaptor sequence) (IDT). The PCR mixture and HFE-7500 fluorinated oil (3M) with 2% (w/w) PEG-PFPE amphiphilic block copolymer surfactant (008-Fluoro-surfactant, Ran Technologies) are loaded into separate 1-mL syringes (BD) and injected at 300 and 500 µL/h, respectively, into a flow-focusing droplet maker using syringe pumps (New Era, catalog no. NE-501), controlled with a custom Python script (https://github.com/AbateLab/Pump-Control-Program). The emulsion is collected in PCR tubes, and the oil underneath the emulsion removed by pipette and replaced with FC-40 fluorinated oil (Sigma-Aldrich, catalog no. 51142-49-5) with 5% (w/w) PEG-PFPE amphiphilic block copolymer surfactant for improved thermal stability. The emulsion is thermal cycled (Bio-Rad, T100) with the following program: 98 °C for 3 min, followed by 40 cycles with 2 °C per second ramp rates of 98 °C for 10s, 62 °C for 20s, and 72 °C for 20s, followed by a hold at 12 °C. Fluorescent DNA staining using 10× SYBR Green I (Thermo Fisher Scientific) in HFE-

7500 oil is used to quantify barcode encapsulation rate under a fluorescent microscope (Life Technologies, catalog no. AMAFD1000).

2.5.3. Water sample collection and filtering

To obtain a natural sample of a microbial community, we collected marine water from Ocean Beach in West San Francisco, California, USA (37°44′55.6″N 122°30′33.6″W). Approximately 2 liters of water is obtained by submerging two 1,000-mL glass bottles below the water surface ~20 m from the shoreline. Samples are placed on ice during transport to the laboratory. 100 mL of the sample is passed through a 40-µm cell strainer (Corning, product no. 352340) to remove large debris, including sand. The sample is loaded into a 0.45 µm vacuum filter (Millipore, catalog no. SCHVU01RE); this filtering step separates microbes, which are captured on the membrane, and viruses, which are discarded in the filtrate. The membrane is extracted from the apparatus using a scalpel and inserted into a 15-mL centrifuge tube, to which 5 mL of PBS is added. The tube is vortexed at high speed for ~2 min to free the bacterial cells from the membrane. Finally, the cell solution is loaded into a 10-mL syringe and passed through a 5-µm syringe filter (Millipore, catalog no. SLSV025LS) to remove remaining large particulate. The marine cells are counted using the same protocol as the liquid cultures.

2.5.4. Cell encapsulation in agarose microgels

To prepare the artificial community for processing through the SiC-seq workflow, the frozen stock of cells (Zymo Research, catalog no. D6300) is thawed gently in a room-temperature water bath. Cell concentration is determined by manual cell counting under a microscope, and diluted to an appropriate concentration for single-cell encapsulation. The calculated volume of cell solution is transferred to a 1.5-mL centrifuge tube (Fisher Scientific) and washed twice in 1 mL PBS. The cells are re-suspended in a 1-mL solution of PBS containing 17% OptiPrep Density Gradient Medium (Sigma-Aldrich), 0.1 mg/mL BSA (Sigma-Aldrich, catalog no. A9418), and 1% (v/v) Pluronic F-68 (Life Technologies). The cell solution is

loaded into a 1-mL syringe and placed on a syringe pump (New Era, catalog no. NE-501). 1 mL of a 3% solution of low gelling temperature agarose (Sigma-Aldrich, catalog no. A9414) and TE buffer (Teknova, catalog no. T0225) is prepared in a 1.5-mL centrifuge tube and heated on a block at 90 °C for approximately 10 min to completely dissolve the agarose powder. The hot agarose is transferred to a 1 mL syringe and placed on a syringe pump. To keep the agarose molten during the microfluidic experiment, a personal space heater is positioned ~5 cm from the agarose syringe and set to run continuously at high heat. HFE-7500 fluorinated oil with 2% (w/w) de-protonated Krytox surfactant (DuPont, catalog no. 157FSH) is loaded into a 3-mL syringe. The cell solution, molten agarose, and oil are injected into the co-flow droplet maker at flow rates of 200, 200, and 400 µL/h, respectively, to form the 1.5% agarose microgels. Approximately 500 µL of droplets are collected in a 15-mL centrifuge tube on ice and incubated for 30 min at 4 °C to ensure complete solidification of the microgels.

2.5.5. Resuspending microgels in aqueous buffer

The droplets are centrifuged at 300g for 1 min to maximize separation of the emulsions from the oil. The oil layer is extracted from the tube using a 5-mL syringe and discarded. Emulsions are broken using 2 mL of a 10% (v/v) solution of perfluoro-octanol (Sigma-Aldrich, catalog no. 370533) in HFE-7500; the emulsions are then mixed by pipetting and centrifuged at 300g for 1 min. The oil is removed from the tube using a syringe, and the droplet breaking step is repeated. Following droplet breaking, 2 mL of hexane containing 1% (v/v) Span 80 (Sigma-Aldrich) is added to the microgels to dissolve any remaining oil, and this solution is mixed and centrifuged at 300g for 1 min. The hexane supernatant is removed from the tube and the hexane addition step is repeated. Finally, the microgels are washed three times in 10 mL of aqueous solution TE buffer containing 0.1% (v/v) Triton X-100 nonionic surfactant (Sigma-Aldrich). The microgels are centrifuged at 1,000g for 2 min and the supernatant aspirated between washes. The washed microgels are stored in 5 mL TE buffer at 4 °C before cell lysis.

2.5.6. Cell lysis in microgels

To lyse the cells in the microgels, the particles are submerged in a solution of 2 mL TE buffer solution containing 10 mM DTT (Teknova), 2.5 mM EDTA (Teknova), and 10 mM NaCl (Sigma-Aldrich). The following quantities of lytic enzymes are also included: 4 U zymolyase (Zymo Research), 10 U lysostaphin (Sigma-Aldrich, catalog no. L7386), 100 U mutanolysin (Sigma-Aldrich, catalog no. M9901), and 40 mg lysozyme (MP Biomedicals, catalog no. 195303). Cell lysis proceeds overnight in a shaking incubator at 37 °C. The turbid lysate mixture is centrifuged at 1,000g for 1 min, the supernatant removed, and 3 mL of a solution containing 0.5% (w/v) lithium dodecyl sulfate (Sigma-Aldrich) and 10 mM EDTA in TE buffer is added, along with 4 U of Proteinase K (NEB) to solubilize cell debris and digest cellular proteins. The solution is incubated at 50 °C on a heating block for 30 min. Following lysis, the microgels are thoroughly washed to ensure complete removal of detergents and other chemical species, which may inhibit downstream molecular biology reactions. The following washes occur in 10-mL volumes with centrifugation magnitudes of 1,000g between additions of wash solutions: one wash with 2% (v/v) Tween 20 in water; one wash in 100% ethanol (Koptec) to denature any remaining Proteinase K; and five washes with 0.02% (v/v) Tween 20 in water.

2.5.7. Tagmentation of genomic DNA in microgels

Using reagents from a Nextera DNA Library Prep Kit (Illumina, catalog no. FC-121-1030), the washed and lysed gels containing high-molecular-weight genomic DNA are simultaneously fragmented and tagged with a common adaptor sequence. Microgels are re-encapsulated into droplets to minimize cross-contamination during the tagmentation step. A solution of 192 µL DI water, 200 µL tagmentation buffer, and 8 µL Nextera enzyme is prepared and loaded into a 1-mL syringe. Microgels and the tagmentation solution are injected into the re-encapsulation device. The re-encapsulated microgels are incubated in a 1.5-mL tube on a heating block at 50 °C for 1 h.

2.5.8. Microfluidic barcoding of encapsulated cells

Tagmented microgel droplets, barcode droplets, and 500 µL of PCR solution containing 1× Invitrogen Platinum Multiplex PCR Master Mix (Thermo Fisher Scientific, catalog no. 4464268), 400 nM primers FL127 (AATGATACGGCGACCACCGAGATCTACACTCGTCGGCAGCGTC, contains P5 adaptor sequence) and FL129 (CAAGCAGAAGACGGCATACGAGATCAGCTGGCGTAATAGCG), 50× dilution of NT buffer from the Nextera XT Kit (0.2% SDS) (Illumina, catalog no. FC-131-1024), 1% (w/v) Tween 20, 1% (w/v) PEG-6000, 2.5 U/µL Bst 2.0 WarmStart DNA Polymerase (NEB, catalog no. M0538S) are each loaded into a 1-mL syringe and injected into the sequential merger device. HFE-7500 fluorinated oil with 2% (w/w) 008-Fluorosurfactant is used as the continuous phase of the emulsion. Merger of the barcode and gel droplet emulsions is achieved using an electrode connected to a cold cathode fluorescent inverter and DC power supply (Mastech). A voltage of 2.0 V at the power supply produces a ~2 kV AC potential at the electrode, which causes touching droplets to merge. The emulsion is collected in a 0.5-mL thin-walled PCR tube (Applied Biosciences), and the HFE-7500 replaced with FC-40 with 5% (w/w) 008-Fluorosurfactant before thermal cycling with the following protocol: 65 °C for 5 min, 95 °C for 2 min, then 30 cycles at 2 °C/s ramp rates of 95 °C for 15s, 60 °C for 1 min, 72 °C for 1 min, and then 72 °C for 5 min with optional 12 °C overnight hold. After thermal cycling, large (coalesced) droplets are removed using a micropipette, and the emulsion is broken by addition of 20 µL of perfluoro-octanol and brief centrifugation in a micro-centrifuge. The upper aqueous phase is collected and the DNA library is purified using a Zymo DNA Clean & Concentrator-5 kit (Zymo Research). The library is size-selected for DNA fragments in the 200- to 600-bp range using Agencourt AMPure XP beads (Beckman Coulter), quantified with a Bioanalyzer 2100 instrument and High Sensitivity DNA chip (Agilent), and sequenced on an Illumina MiSeq using a custom index primer (GCCCACGAGACGTGTCGGGGCTGGCTTA).

2.5.9. Generating the SiC-Reads database

Raw reads from the MiSeq-generated FASTQ files are filtered by quality and grouped by barcode sequence using the Python script barcodeCleanup.py. A given read is discarded if more than 20% of its bases have a Q-score less than Q20, and all reads associated with a barcode containing less than 50 reads are discarded. This step ensures that all barcode groups, representing single cells, contain a sufficient number of high-quality reads. The resulting reads are exported to a table in a SQLite database with fields containing the barcode sequence, barcode group size, a unique read ID number, and read sequence. When the reference genomes are known, as in the case of the synthetic cell population experiment, the reads are aligned using bowtie2 v2.2.9 with default settings and the SQLite table is updated with relevant alignment information for each read. For environmental samples, the reads are classified by taxonomy using Kraken v0.10.5 with "–quick–min-hits 2" options set, and the output is exported to the SQLite database. Where noted, an alternative taxonomic classifier, metaphlan2 v2.6.0, was used with the default marker information file and taxonomic level set to "species only." krakenAnalysis.py assigns taxonomic identities from the Kraken database to barcode groups by a majority rule, in which a barcode group is classified according to the most common taxonomic label among its classifiable reads. Barcode group purity is calculated from reference alignment data or phylogenetic labels using the script purity.py.

2.5.10. In silico *cytometry*

Reads from the SiC-Reads database are aligned, using bowtie2 v2.2.9 with –very-sensitive and –end-to-end settings to reference sequences of interest. The antibiotic resistance database was obtained from[36], virulence factor database obtained from core virulence factor genes at the virulence factor database (VFDB)[37], and the phage sequence database obtained from Phage genome database accessed on May 2016 at http://www.ebi.ac.uk/genomes/phage.html. Mapping reads are then filtered for MapQ > 2 in

order to remove ambiguously mapping reads. Barcode groups containing reads that map to the databases are annotated as containing the target sequence and are exported for further analysis if they are taxonomically classified with purity >0.8. To generate the heatmap for transduction potential, all reads associated with a phage and a Kraken-classified barcode group were extracted and grouped according to phage type. Duplicate and near-duplicate reads were removed. The heatmap intensities were calculated as follows: for a given pair of bacterial hosts, the total number of host-phage-host connections in the database were counted. To normalize the data by host abundance, this number was divided by the total number of barcode groups associated with the two hosts.

2.5.11. Calculating the virulence factor ratios

The virulence factor ratios calculations in **Figure 2.14b** of the main text was reproduced using reference genomes for the genera shown in the figure. The complete genomes of all species associated with these 12 genera were downloaded from the RefSeq database using the Perl script ncbiDownloader.pl. Genomes were pooled into FASTA files labeled by genus. From these reference files, a Python script (bargroupGenerator.py) generated simulated barcode groups of 200 reads per group, with each single-end read 150-bp long. The number of simulated barcode groups generated for a given genus was equal to the number of barcode groups identified for this genus in the San Francisco Coast water sample. The simulated barcode group reads were then aligned to the original virulence factor database using bowtie2 v2.2.9 in 'local' alignment mode with default sensitivity settings. Unaligned sequences were removed using Samtools v1.3.1 (samtools view -b -F).

2.5.12. Generating the antibiotic-resistance network with reference genomes

An antibiotic resistance graph in was generated using references for the six genomes most commonly associated with antibiotic resistance in the SiC-Reads database of the San Francisco Coast water microbial community. The following genomes (with accession numbers)

were downloaded from the NCBI RefSeq repository: *Alteromonas macleodii* ATCC 27126 (CP003841.1), *Bacillus subtilis subsp. spizizenii* strain NRS 231 (CP010434.1), *Delftia acidovorans* SPH-1 (CP000884.1), *Enterobacter cloacae subsp. cloacae* ATCC 13047 (CP001918.1), *Neisseria meningitidis* MC58 (AE002098.2), *Propionibacterium acnes* KPA171202 (AE017283.1). These genomes were combined into a single FASTA file and passed to a short-read simulator, wgsim v0.3.2, which generated 10 million single-end reads of 70 bp each with a base error rate of 0. These reads were aligned to the antibiotic-resistance gene reference using bowtie2 in 'local' mode with default sensitivity settings. All unaligned sequences were removed using Samtools (samtools view -b -F). The aligned sequences in SAM format were imported into Cytoscape v3.4.0, and the network shown was generated using the reference genus and antibiotic-resistance genes as the network targets and sources, respectively. The darkness of the graph's edges scale linearly with the total number of connections in the data, where darker lines have a greater number of associations.

2.5.13. Characterizing diffusion of genomic DNA fragments in agarose microgels

A microgel sample containing encapsulated, lysed bacteria was stained with SYBR Green I and observed under a fluorescent microscope before and after tagmentation at various time points. In another experiment, the concentration of DNA in the supernatant and contents of the gels was measured in a sample incubated at room temperature. After 2 d at room temperature, the beads were pelleted by centrifugation; the DNA was extracted from the beads and from the supernatant, using a DNA gel extraction kit and a DNA clean-up and concentrator kit (Zymo research D4001T); the extracted DNA was quantified using the Qubit dsDNA high sensitivity assay and Bioanalyzer High Sensitivity dsDNA chip. As an additional experiment, microgels were incubated at 55 °C with and without tagmentation enzyme to demonstrate the corresponding change in genomic DNA fragment size distribution before and after tagmentation.

2.5.14. Cell culture and counting

To generate an additional artificial community with which to validate the SiC-seq workflow, liquid cultures of *Staphylococcus epidermidis*, *Saccharomyces cerevisiae* (strain S288c), and *Bacillus subtilis* (strain 168) were grown overnight in a shaking incubator. The following culture conditions were used: *Staphylococcus epidermidis* and *Bacillus subtilis* were grown in 3-mL LB broth at 37 °C; *Saccharomyces cerevisiae* was grown in 3 mL YPD broth at 30 °C. Cell concentration is determined by manually counting serial dilutions of the liquid culture on plastic slides (Thermo Fisher Scientific, catalog no. C10228) using a microscope. The cultures were kept at 4 °C before being used in the microfluidic experiment.

Chapter 3: Direct quantification of *EGFR* variant allele frequency in cell-free DNA using a microfluidic-free digital droplet PCR assay

3.1. Abstract

Analysis of liquid biopsy samples is a promising diagnostic intervention for noninvasive detection and monitoring of cancer genotypes. However, current methods used to assess mutation status are either costly, in the case of next-generation sequencing-based assays, or lacking in sensitivity, in the case of bulk quantitative PCR measurements. Digital droplet PCR (ddPCR) is at once a sensitive and low-cost method for detecting rare cancer mutations and measuring their variant allele frequency. In this chapter, we describe a method for conducting ddPCR assays without microfluidics in a process called “particle-templated emulsification” (PTE). Using hydrogel particles and a standard benchtop vortexer to rapidly emulsify large volumes, the method forgoes the specialized instrumentation required for conventional ddPCR assays and is capable of high experimental throughput. To assess the quantitative performance of the method, we apply PTE ddPCR to analysis of variant allele frequency in *EGFR*, a commonly mutated gene in lung adenocarcinomas.

3.2. Introduction

Lung cancer has the highest mortality rate of all cancers, accounting for 19.4% of all cancer-related deaths worldwide[38,39]. Lung adenocarcinoma (ADC) in non-small-cell lung cancer (NSCLC) accounts for most primary lung cancers[38]. Within the ADC varieties, mutation status in the epidermal growth factor receptor (EGFR) gene is a strong predictor of therapeutic response and metastatic stage[40]. However, invasive tissue biopsies are difficult and impractical for routine clinical management and detection of new cases of cancer. Cell-free DNA (cfDNA), DNA which is released into plasma from tumor tissues or circulating tumor cells, harbors the cancer mutations and is accessible using a simple liquid biopsy. A technical challenge associated with cfDNA-based diagnostics, however, is the high background of host DNA. The variant allele

frequency (VAF) of the cancer allele compared to wildtype can be as low as 0.1%, necessitating sensitive assays[41].

Next-generation sequencing offers the highest sensitivity of existing methods but is costly and has a longer turnaround time, delaying clinical action[42]. Quantitative PCR is the most widely used clinical diagnostic method for cfDNA, but has a high limit of detection (2–5% minimum variant allele frequency for the commercial Roche Cobas kit) and does not allow for direct determination of allelic frequency of cancer mutations[43]. Digital droplet PCR has higher sensitivity and can directly quantify allelic frequency of cancer mutations without need of a standard curve because it provides an absolute count of all variants[41]. However, existing digital droplet PCR (ddPCR) platforms rely on proprietary droplet analysis hardware, which is expensive. Furthermore, microfluidic devices are required to process the sample into droplets, which are prone to clogging, limit the number of samples that can be analyzed, and require transfer of reagents between tubes and microfluidic components, increasing the chances for sample contamination. A digital droplet assay without complex hardware and sample processing bottlenecks would enable rapid and quantitative profiling of cfDNA for making clinical decisions related to lung cancers.

We describe a method to detect and quantify the variant allele frequency of an *EGFR* mutation using a digital droplet PCR assay that uses no microfluidics or specialized equipment. The approach is based on particle-templated emulsification (PTE) which partitions reactions into monodispersed droplets via simple vortexing of the sample container. Reagents do not need to be transferred between components, and the approach can emulsify large numbers of samples in parallel stored in microplate arrays. Moreover, the time to emulsify the sample is independent of the sample volume and takes ~ 30 s. As we demonstrate, PTE droplets can be used to detect and quantify *EGFR* mutants in a sample. Moreover, by multiplexing the reaction and using two hydrolysis probes labeled with dyes of different colors, we can identify wildtype or mutant alleles in droplets and quantify variant allele frequency (down to 1%) with fluorescence imaging data.

Our method should enable more sensitive ddPCR by being compatible with high-volume clinical sample processing and by eliminating the need for specialized droplet analyzers.

3.3. Discussion

In ddPCR, a sample containing DNA to be quantified is encapsulated into millions of picoliter-volume droplets, such that most partitions are devoid of the target and a small fraction contain it. At such limiting dilution, droplet occupancy by the targets follows a Poisson distribution, allowing the measured fraction of positive droplets to be converted into a concentration measurement. Digital quantification affords significant advantages over real time measurement of amplification rates as in qPCR, because the sensitivity increases as the number of partitions is increased, and there is no need for a standard curve by which to interpolate starting concentration from amplification kinetics, as in qPCR. Digital droplet PCR thus affords higher sensitivity and absolute molecule counts, making it valuable for applications requiring robust and accurate quantitation of targets, such as in the clinic. Consequently, ddPCR has been used to measure variant allele frequencies of DNA samples harboring rare mutations, such as in cell-free DNA (cfDNA) from plasma of patients with cancer. The cfDNA originating from circulating tumor cells and tissue contains mutations which can provide information about the metastatic potential and therapeutic susceptibility of a given condition[40,41].

In lung adenocarcinomas, mutations in the *EGFR* gene are common but difficult to detect in a liquid biopsy due to the high background of DNA from other host cells without the mutation[44]. Here, we describe a digital PCR method employing particle-templated emulsification (PTE) for quantification of a specific *EGFR* mutation without the use of complex microfluidic devices or well plates. Much like conventional digital PCR in droplets, the water-in-oil emulsions serve as reaction partitions and allow digital amplification of target template. However, rather than using a microfluidic device to generate the droplets, PTE uses polyacrylamide microgels to engulf the sample into the partitions. A powerful advantage of PTE is that all of the

emulsification occurs in the original sample tube; thus, the sample need not be transferred between reservoirs, tubing, and microchannels, as in microfluidic devices, making the approach simple and robust to contamination. Moreover, while in microfluidic emulsification the time to process the sample scales with the sample volume, with PTE emulsification takes ~ 30 s, for samples ranging from tens of microliters to milliliters. These properties make the approach scalable in the number of samples processed and in the volume processed per sample.

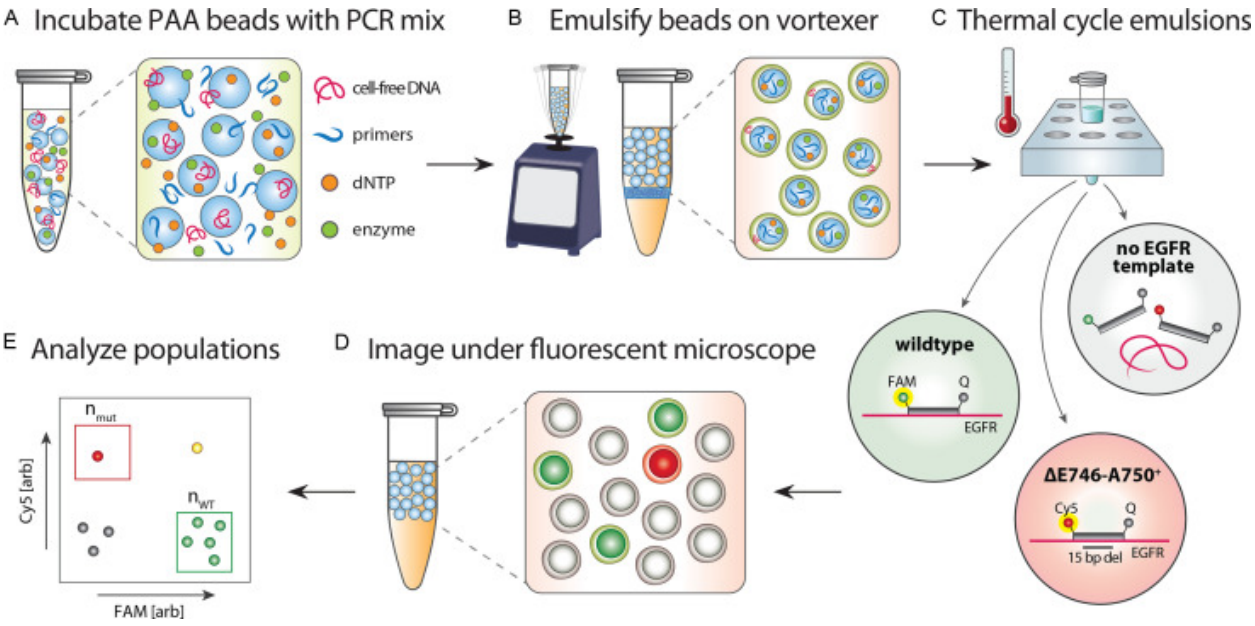

Figure 3.1: Digital PCR workflow using particle-templated emulsification.
(a) Polyacrylamide beads are soaked in PCR mix, allowing amplification components to diffuse throughout the bead volume. **(b)** Beads are emulsified using a simple laboratory vortexer which creates the necessary hydrodynamic shear for efficient partitioning. **(c)** The contents of the emulsions are PCR-amplified in a conventional thermal cycler. In partitions containing the target nucleic acid sequence, the polymerase cleaves the hydrolysis probes and emits a fluorescent signal. **(d)** Droplets are imaged under a microscope to capture fluorescence intensity values within each droplet. **(e)** Populations of droplets are analyzed using image processing tools. Counts of mutant and wildtype droplets are used to calculate the variant allele frequency.

The workflow begins by adding the microgels to the sample, which contains the target DNA template and amplification reagents. Depending on the porosity of the hydrogel particles, some components of the sample, such as small molecules, will diffuse into the hydrogels; larger molecules, like macromolecular DNA, will remain in the voids between hydrogels, as illustrated

in **Figure 3.1a**. Upon addition of carrier oil and surfactant and agitation by vortexing, the sample is emulsified into droplets. Each droplet consists of a single bead surrounded by a thin shell of amplification reagent encapsulated in oil (**Figure 3.1b**). Because the particles are monodisperse, and provided the agitation is somewhat controlled, the sample and droplet volumes are monodisperse. This yields an emulsion that is equivalent to microfluidic emulsions commonly used for ddPCR, without microfluidics. Thus, we thermal cycle the emulsion, allowing amplification of the mutant and wildtype *EGFR* allele using common primers and dual-color hydrolysis probes (**Figure 3.1c**). In this experiment, we target a common mutation in lung adenocarcinomas, a 15 bp deletion on exon 19 of the *EGFR* gene, *EGFR* ΔE746-A750. The mutation is associated with increased resistance to EGFR inhibitors used in the treatment of non-small-cell lung cancers[38].

To detect the positive droplets, we use fluorescence microscopy, processing tens of thousands of droplets (**Figure 3.1d**). Analysis of fluorescence measurements using image processing software shows populations of droplets which contain no *EGFR* template, wildtype *EGFR*, mutant *EGFR*, and double positives (**Figure 3.1e**). The relative counts of the wildtype and mutant populations provide a measurement of variant allele frequency. To demonstrate the process, we use the approach to analyze two human cfDNA standards containing the mutant ΔE746-A750 allele at frequencies of 5% and 1% versus wildtype *EGFR*. In this experiment, the hydrogels are ~ 100 μm in diameter and monodisperse (**Figure 3.2a**). Following addition of the DNA template and amplification reagents, vortexing yields droplets with single particles surrounded by sample (**Figure 3.2b**). We thermal cycle and image the droplets for wildtype and mutant *EGFR* variant frequencies. A merged image of the fluorescein (wildtype) and Cy5 (mutant) channels shows discrete signal from droplets containing an *EGFR* variant (**Figure 3.2c**). We use image analysis to measure the fluorescence of the droplets (**Figure 3.2d**) and find that the measured allelic frequencies agree with the input concentration values of 5% and 1%. These results demonstrate that PTE allows rapid and facile measurement of *EGFR* variant

alleles with accuracy comparable to ddPCR, but without the use of microfluidics and in a low-cost and scalable format.

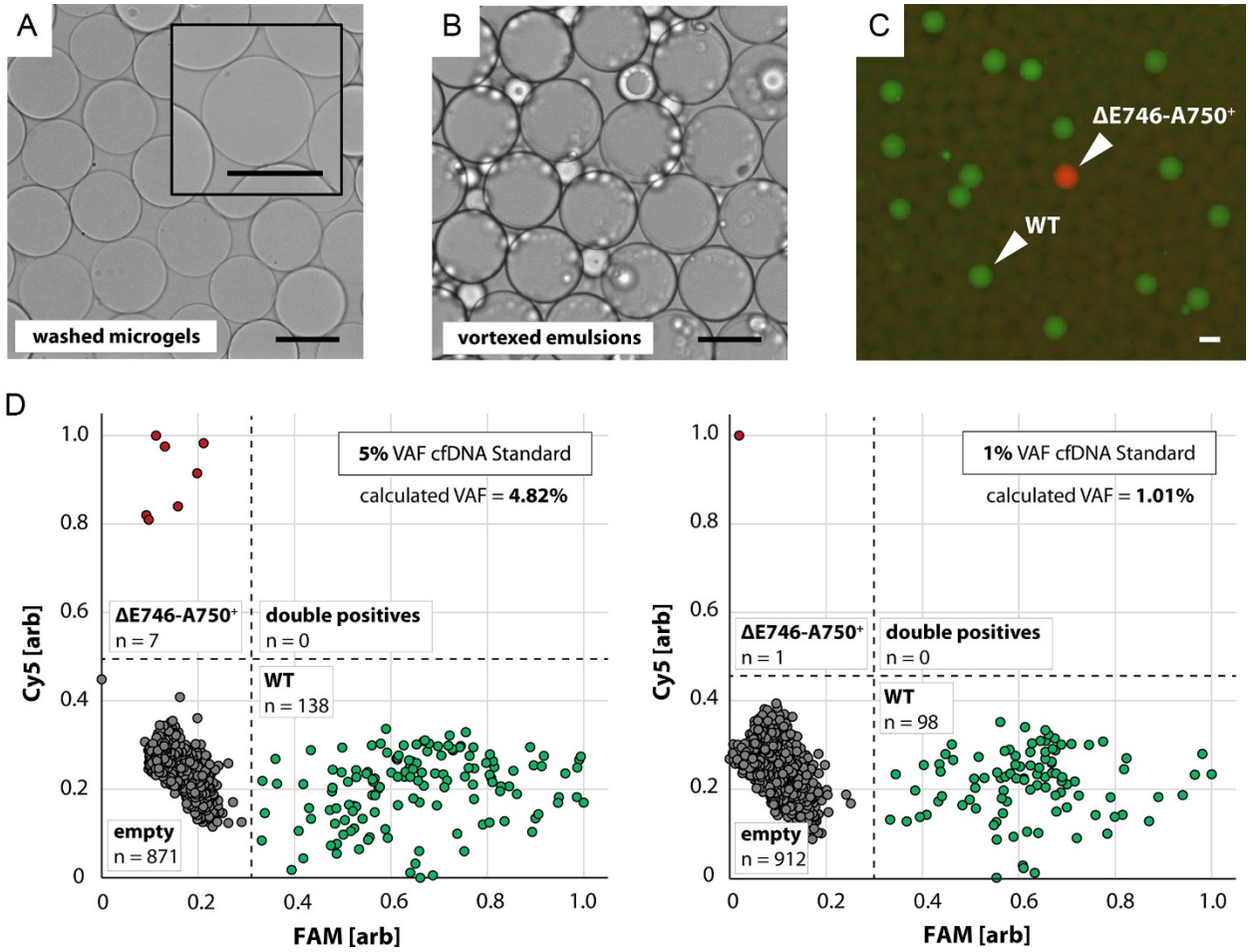

Figure 3.2: Variant allele frequency analysis of cfDNA samples using a hydrogel-partitioned digital PCR assay.
(a) After washing, the microgels appear translucent when suspended in an aqueous buffer. **(b)** Vortexing the microgels with fluorinated oil and PCR buffer produces monodisperse emulsions each containing a single microgel surrounded by a shell of amplification reagent. **(c)** Representative fluorescence micrograph of thermal cycled cfDNA emulsions containing the mutant *EGFR* allele at 5% relative frequency. **(d)** Scatter plots of normalized droplet fluorescence values for cfDNA samples containing the *EGFR* mutation at 5% and 1% allelic frequencies. Scale bars are 100 μm in all imaging panels.

3.4. Methods

3.4.1. Choice of polyacrylamide microgels for PTE

The PTE-based digital droplet assay requires monodisperse hydrogel particles for efficient emulsification of the DNA template and amplification reagents. A particle suspension containing microgels varying widely in size produces inaccurate quantification measurements, as the volume of the sampling container is non-uniform. In this chapter, we describe the process for generation of hydrogel particles using a microfluidic dropmaker. This method reliably produces monodisperse hydrogel suspensions for PTE, as well as gives the user control over the size of the particles.

For users without microfluidic expertise or those lacking access to microfluidic equipment, an alternative option is to use commercially-available hydrogel microspheres (e.g., Bio-Rad Bio-Gel P-60 Gel, catalog no. 150-4164, or equivalent). These are typically marketed as components for purification columns and are therefore free of contaminants and common PCR inhibitors. Prior to use in the assay, users should wash the hydrogels thoroughly after rehydrating and remove very small and large particles by passing the suspension through a cell strainer.

3.4.2. Microfluidic preparation of hydrogel particles

The following protocol uses a microfluidic dropmaker to generate microgels for PTE. The dropmaker was fabricated using standard soft lithography techniques, detailed protocols for which can be found in prior publications[2,45]. For this digital droplet assay, the photoresist master mold should be fabricated with a feature height of 70 μm. Additionally, the oil and aqueous reagent channels at the dropmaking junction should have widths of 70 μm each, to produce droplets with an approximate final diameter of 70 μm.

3.4.3. Generate the hydrogel particles using a microfluidic dropmaker

1. Prepare 1 mL of acrylamide precursor solution according to the recipe shown in **Table 3.1**.

Table 3.1: Recipe for the acrylamide precursor solution.

Component	**Final Concentration**
Tris buffer, pH 7.5	10 mM
Ethylenediaminetetraacetic acid (EDTA)	1 mM
Sodium chloride	15 mM
Acrylamide	6.20% (w/v)
N,N'-Methylenebis(acrylamide)	0.18% (w/v)
Ammonium persulfate	0.30% (w/v)

2. Prepare the oil phase by adding 1% (v/v) tetramethylethylenediamine (Sigma-Aldrich) to 2 mL fluorinated oil (HFE-7500, 3M) containing 2% (w/w) PFPE-PEG surfactant (008-Fluorosurfactant, Ran Biotechnologies).
3. Load the acrylamide and oil solutions into 3-mL syringes (BD).
4. Setup two syringe pumps (New Era, NE-501) to drive the polyacrylamide precursor solution and fluorinated carrier oil into the dropmaking device. Connect all syringes to their respective inlets and connect a piece of tubing from the outlet to a 15-mL collection tube.
5. Generate droplets using flow rates of approximately 500 and 1000 µL/h for the acrylamide and oil phases, respectively. Pumps are controlled by a custom Python script (available at https://github.com/AbateLab/Pump-Control-Program). During dropmaking, users should periodically observe droplets under a microscope to ensure they are monodisperse with an average diameter of approximately 70 µm.
6. After dropmaking, incubate the collected droplets on a heat block at 65 °C for 1 h to polymerize, then, proceed to the wash steps.

3.4.4. Wash the hydrogel droplets

1. Prepare 100 mL of hydrogel wash buffer using the recipe shown in **Table 3.2**.

Table 3.2: Recipe for the hydrogel wash buffer.

Component	**Final Concentration**
Tris buffer, pH 8.0	10 mM
Ethylenediaminetetraacetic acid (EDTA)	10 mM
Sodium chloride	137 mM
Potassium chloride	2.7 mM
Triton X-100	0.1% (v/v)

2. Remove excess oil from the bottom of the emulsion tube using a gel-loading pipet tip.
3. Add an equal volume of perfluoro-1-octanol (Sigma-Aldrich) to the emulsions to dissolve their surfactant layer. Pipet up and down thoroughly to completely coat the emulsions.
4. Incubate the emulsions 2 min at room temperature.
5. Centrifuge the emulsions at 3000 × g for 1 min. After centrifugation, the hydrogels will be free of their surfactant layer and appear translucent.
6. Remove excess perfluoro-1-octanol using a gel-loading pipet tip.
7. Add 10 mL of wash buffer to the hydrogel tube. Vortex 10 s to resuspend hydrogels.
8. Centrifuge the hydrogel suspension at 3000 × g for 3 min.
9. Aspirate the supernatant, taking care not to disrupt the translucent hydrogel layer at the bottom of the tube.
10. Repeat steps 8 and 9 two additional times to completely remove all oil from the hydrogel suspension.
11. Resuspend the beads in 5 mL of wash buffer and store at 4 °C until use in the PTE experiment.

Users should examine the hydrogels under a microscope after washing to ensure complete removal of oil and other debris. Particles should appear translucent, spherical, and monodisperse, similar to those shown in **Figure 3.2a**. If debris or clumps are observed, the

hydrogel suspension can be filtered through a 100 μm cell strainer. Once prepared, hydrogels can be stored in wash buffer for a year or longer at 4 °C without any noticeable loss of particle integrity.

3.4.5. Digital droplet PCR assay using particle-templated emulsification

After hydrogel particle generation, PCR template and reagents are emulsified in a simple, microfluidic-free protocol. No further specialized equipment is required for performing the PTE-based digital droplet assay. The particles generated in the preceding steps provide sufficient hydrogels for tens of reactions (depending on reaction volume), and a typical ddPCR assay using PTE requires only several minutes of hands-on time to perform.

In the following experiment, a cell-free DNA standard sourced from a commercial vendor (Multiplex I cfDNA Reference Standard Set, Horizon Discovery) is used to assess the quantification performance of the *EGFR* assay. The assay is readily extensible to real samples of human sera in which the relative frequency of the variant allele is unknown. In all cases, some optimization on behalf of the user is required to achieve an encapsulation rate of the target template within an optimal range for quantification. To limit the rate of double encapsulation events, an overall encapsulation rate of one template copy for every 10 droplets is optimal. A rough estimate of the amount of DNA template required per reaction can be calculated using the volume of the reaction partitions, accounting for an aqueous shell thickness of 5 μm around each hydrogel particle. Users are advised to first perform the assay with calculated amounts of template and adjust the input accordingly based on observed encapsulation rate after amplification.

3.4.6. Prepare the reaction components

1. Prepare ~ 200 μL of digital PCR solution by combining the components shown in **Table 3.3**. Pipet up and down thoroughly to mix.

Table 3.3: Recipe for PCR amplification mix (primers shown for the *EGFR* ΔE746-A750 assay).

Component (Concentration)	**Volume**	**Final Conc. (approx.)**
Platinum Multiplex PCR Master Mix (2X)	200 µL	1X
FWD primer (100 µM) 5' GGATCCCAGAAGGTGAGAAAG 3'	1.6 µL	0.4 µM
REV primer (100 µM) 5' CAGCAAAGCAGAAACTCACATC 3'	1.6 µL	0.4 µM
WT probe (100 µM) 5' /56-FAM/CGCTATCAA/ZEN/GGAATTAAGAGAAGCAACATCTCC/3IABkFQ/ 3'	0.8 µL	0.2 µM
MUT probe (100 µM) 5' /5Cy5/CGTCGCTAT/TAO/CAAAACATCTCCGAAAGC/3IAbRQSp/ 3'	0.8 µL	0.2 µM
Triton X-100	4 µL	2% v/v

2. For each individual reaction, prepare a 50 µL aliquot of the digital PCR solution in a 1.5-mL microcentrifuge tube, then add the predetermined volume of DNA template. To limit dilution of amplification reagents, the volume of added DNA template should not exceed 5 µL. Mix thoroughly by pipetting.
3. Prepare the polyacrylamide gels by centrifuging the stock suspension at 3000 × g for 3 min. Remove the wash buffer supernatant by aspiration.

3.4.7. Generate the particle-templated emulsions

1. Transfer 50 µL of polyacrylamide gels to each microcentrifuge tube containing PCR mix using a wide-bore pipette tip. Pipet up and down thoroughly to mix.
2. Incubate each reaction for 15 min at room temperature on a rotating tube mixer.
3. Spin down each reaction at 6000 × g for 1 min. Carefully aspirate the supernatant, leaving only the hydrogel pellet.
4. After removing the supernatant, add 100 µL of HFE-7500 fluorinated oil containing 2% (w/w) PFPE-PEG surfactant to the bottom of the microcentrifuge tube under the polyacrylamide gel pellet.

5. Vortex each tube individually at maximum power for 30 s. Slowly decrease the vortex speed over the course of 10 s to minimize the amount of emulsions on the tube walls. Allow each reaction to sit for 5 min at room temperature to allow the emulsions to settle.
6. Using a gel loading pipet tip, remove the bottom oil phase from the tube along with the lower layer of emulsions. The emulsions will settle into two distinct layers, with the larger hydrogel-containing emulsions at the top. Rotating the microcentrifuge tube in front of a light source can help to identify this distinct layering.
7. Finally, transfer 30 µL of emulsions from each reaction into labeled PCR tubes using a pipet. Add 40 µL of FC-40 fluorinated oil (Sigma-Aldrich) containing 5% (w/w) PFPE-PEG surfactant to each PCR tube. At this stage, the emulsions should each contain a single hydrogel and appear similar to those shown in **Figure 3.2b**.
8. Thermal cycle the tubes using the following protocol: 95 °C for 2 min, 40 cycles of (95 °C for 30 s, 58 °C for 30 s, 72 °C for 30 s), 4 °C hold.

3.4.8. Image the thermal cycled emulsions

1. To image each sample, carefully collect 8 µL of the oil layer and 2 µL of droplets from each tube using a pipet and transfer into a plastic imaging slide (Countess cell counting chamber slides, Thermo Fisher, catalog no. C10228, or equivalent).
2. Focus the image on the brightfield channel. The vast majority (> 95%) of droplets should contain a single hydrogel. A large number of coalesced droplets may be an indication of improper handling and/or thermal instability of the surfactant layer.
3. The wildtype probe is imaged on the GFP channel (470/22 nm Ex, 510/42 nm Em) while the mutant allele probe is imaged on the Cy5 channel (628/40 nm Ex, 692/40 nm Em). Verify that the fluorescent signal from droplets on both channels is discrete, as shown in **Figure 3.2c**.

4. For each field of view, capture brightfield, GFP, and Cy5 channel images. Repeat for the entire imaging slide. Depending on the microscope hardware, this process may be automated.

3.4.9. Data analysis and calculation of variant allele frequency

The image-processing software Fiji is recommended for analysis of droplet images[46]. A brief overview of the image processing workflow is provided in this section. For a detailed overview of software functionality, users should consult the Fiji documentation.

1. To image each sample, carefully collect 8 µL of the oil layer and 2 µL of droplets from each tube using
2. For each field of view, import the brightfield, GFP, and Cy5 channel images into Fiji.
3. Apply a threshold to the brightfield image such that the dark droplet boundaries are added to the background.
4. Perform a particle analysis to isolate the circular inner phase of each droplet. Adjust particle circularity and size to select only for droplets containing a single hydrogel.
5. Apply the particle overlay from the brightfield channel to the GFP and Cy5 channels. Measure the mean pixel intensity within all particles and export to graphing software (for example, Microsoft Excel).
6. Repeat for multiple fields of view, until a sufficient number of droplets have been analyzed.
7. In the graphing software, normalize the pixel intensities for each channel using the maximum and minimum values in each channel.
8. Plot the normalized channel intensities on a 2-D scatter plot. Each quadrant should contain a distinct droplet population:
    - **GFP$^-$/Cy5$^-$**: droplets not containing a copy of the *EGFR* locus, either wildtype or mutant.

- **GFP$^+$/Cy5$^-$**: droplets containing a copy of the wildtype *EGFR* allele.
- **GFP$^-$/Cy5$^+$**: droplets containing a copy of the mutant *EGFR* allele.
- **GFP$^+$/Cy5$^+$**: droplets containing copies of both the wildtype and mutant *EGFR* alleles.

The scatter plots in **Figure 3.2d** show the results for a single field of view from cfDNA standards containing 5% and 1% variant allele frequencies (VAF), where VAF is defined as

$$\mathrm{VAF} = \frac{\text{number of droplets containing the mutant allele}}{\text{total number of droplets containing an EGFR allele}}$$

In each case, the calculated VAF is close to the manufacturer's specifications, demonstrating the quantification performance of the ddPCR assay using PTE. No double positive (GFP$^+$/Cy5$^+$) droplets were observed for either sample.

3.5. Conclusion

We have demonstrated the performance of a digital droplet PCR assay capable of detecting variant allele frequencies as low as 1% without the use of sophisticated microfluidic equipment. This surpasses the quantitative limit of bulk qPCR methods, does not require experimental standards, and is low-cost and rapid. Experimental throughput, important for effective clinical implementation, is a notable advantage of this method over traditional digital PCR assays using dedicated microfluidic dropmaking instruments. In future studies, the degree of multiplexing can be pushed further by optimizing fluorescent dye color and concentration to enable detection of a larger number of genotypes. Furthermore, since the sensitivity of ddPCR increases as the sample is partitioned into more droplets, flow cytometric analysis of the droplets affords a route toward even greater sensitivity.

Chapter 4: Joint profiling of proteins and DNA in single cells reveals extensive proteogenomic decoupling in leukemia

4.1. Abstract

Current leukemia therapies target cancer cells with specific phenotypes or genotypes, but this assumes that either genomic mutations or immunophenotypes alone serve as faithful proxies for treatment response. Moreover, the heterogeneity inherent to all cancers, including leukemias, makes direct mapping of genotype-phenotype relationships challenging. Here, we present a method to genotype and phenotype single cells at high throughput, allowing direct characterization of proteogenomic states on tens of thousands of cancer cells rapidly and cost-efficiently. Using this approach, we analyze the disease of three leukemia patients over multiple treatment timepoints and recurrences. We observe complex genotype-phenotype dynamics and extensive decoupling of the relationships over disease progression and response to therapy, illustrating the subtlety of the disease process and the inability to use genotypes as direct proxies for phenotypes. Our technology has enabled the first rigorous test of the prevailing paradigm that treatment of a disease phenotype is equivalent to treatment of its underlying genotype. More broadly, our results highlight the power of single-cell multiomic measurements to resolve complex biology in heterogeneous populations and illustrate how this information can be used to inform treatment. We thus expect that our methodology will find broad application to study proteogenomic tumor landscapes across cancers and will support the next generation of immunotherapy.

4.2. Introduction

Acute myeloid leukemia (AML) is an aggressive hematologic malignancy prone to relapse that often manifests as a polyclonal ensemble of cells with distinctive genotypes but diverse immunophenotypes[47,48]. Because of this disparity, it is difficult to directly link genotypes to immunophenotypes beyond circumstantial evidence from epidemiologic studies. Moreover,

while AML blasts often exhibit immunophenotypes distinct from normal cells, with some surface markers even serving as therapeutic targets[49], genotypes are the strongest prognostic factors, suggesting a weak correspondence between these domains[50,51]. Cellular heterogeneity is an intrinsic aspect of essentially all cancers, including leukemias. Because cancer cells are heterogeneous in genotype and phenotype, single-cell analysis provides a powerful tool for characterizing this complexity and thereby advancing our understanding of different cancers. The value of single-cell analysis is its ability to correlate co-occurrence of different features in individual cells, with high-throughput technologies permitting analysis of thousands of cells to generate rich and intricate feature maps. For example, single-cell genotyping of AML-relevant loci has revealed co-occurrence of mutations and mapping of the clonal relationships between blasts[52–55]. These studies, however, have yet to map DNA genotypes and phenotypes in the same cells, precluding direct linkage of phenotypes to the genetic mutations that drive them.

To obtain simultaneous genotype and immunophenotype information, single cells can be sorted based on multi-parametric antibody analysis, and sequenced. While severely limited in throughput, these studies have uncovered important insights into the genetics of AML, identifying relevant aberrations such as single nucleotide polymorphisms (SNPs) and gene fusions[56]. Single-cell RNA sequencing (scRNA-seq) has emerged as a potentially valuable approach for genotype-phenotype linkage because it is cost effective and scalable[53,57–59]. The mRNA sequences provide genotype information[59,60] while their counts relate phenotype[8,9,61–63]. Moreover, modern approaches are extremely high throughput, allowing characterization of thousands of cells. Nevertheless, genotyping from mRNA remains a challenging and error-prone procedure that, even in the best case, provides incomplete information. For example, stochastic gene expression, biological biases[64], and limited coverage of essential genes combine to make assigning a genotype more difficult than can be achieved by direct analysis of DNA. Moreover, since RNA methods analyze only the expressed portion of the genome, mutations in intronic and other non-transcribed elements, like transcription factor binding sites,

are omitted[65,66]. Thus, while several technologies have highlighted the importance of high-throughput single cell genotype-phenotype measurements, none provide the scalability and precision for comprehensive and accurate mapping of these clinically valuable biomarkers.

In this paper, we describe DAb-seq, a novel approach for joint profiling of DNA and surface proteins in single cells at high throughput. While existing methods attempt to obtain this information from the transcriptome alone, ours directly characterizes DNA for genotype and surface proteins for phenotype – both the gold standards in AML for these annotations. Our approach is thus complementary to scRNA-seq methods and, as we show, provides novel and important information for characterizing the disease. To illustrate the power of DAb-seq, we characterize the immunophenotypic and genotypic diversity underpinning AML in three patients at multiple timepoints, exploiting its throughput to characterize 50 DNA targets and 23 hematopoietic markers in a total of 54,717 cells. This analysis allows tracking of proteogenomic dynamics for multiple patients over multiple treatments and recurrences. We identify extensive genotype-phenotype decoupling, observing immunophenotypic heterogeneity among cells with a shared pathogenic mutation and genotypically diverse cells with a convergent malignant immunophenotype. These findings indicate substantial variability of blast fate upon treatment in AML, and that independent phenotype or genotype measurements do not adequately capture the proteogenomic heterogeneity. More broadly, our work demonstrates how single-cell technologies can inform the diagnosis and treatment of AML by elucidating the complex interplay between DNA mutations and their effects on protein expression.

4.3. Results

4.3.1. Combined single-cell DNA sequencing and antibody profiling (DAb-seq) robustly delineates single-cell genotypes and immunophenotypic diversity

The commercially available Mission Bio Tapestri supports highly multiplexed targeted sequencing of thousands of single cells and is being used across cancers for genotype and

lineage mapping[54]. While the instrument runs a flexible workflow, it does not natively support Abseq, a separate method we developed[67] that allows characterization of single-cell surface proteins by sequencing, and is analogous to flow cytometry in its ability to provide immunophenotype information. Thus, our major technical innovation is to adapt Tapestri to enable simultaneous DNA and Abseq analysis. As in our published Abseq approach, DAb-seq begins with immunostaining of a cell suspension using a mixture of antibody-oligo conjugates (**Figure 4.1a**). Each antibody is associated with a known oligo tag; thus, when cells are stained with a pool of tagged antibodies, each cell is bound with a combination of antibodies and their tags based on surface protein profile. To characterize the profile, the tags must be sequenced and counted which, in flow cytometry, is analogous to measuring fluorescence of the dyes associated with each antibody, except that photon counting is replaced with tag counting.

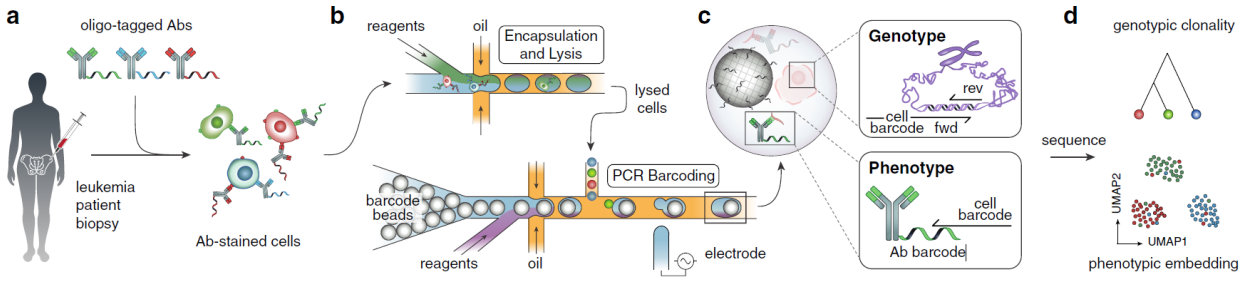

Figure 4.1: The DAb-seq workflow.
(a) Bone marrow aspirates of patients with AML contain healthy and malignant cells that exhibit diverse genotypes and immunophenotypes. These cells are stained with antibodies labeled with DNA tags. **(b)** Stained cells are paired and encapsulated with a barcode bead on a Mission Bio Tapestri instrument. **(c)** In each droplet, a PCR labels antibody tags and genomic DNA targets simultaneously with a unique cell index. **(d)** Sequencing the barcoded amplicons and antibody tags yields coupled single-cell immunophenotype and genotype data for thousands of cells.

The stained cells are processed through a modified Tapestri workflow to amplify and barcode genomic targets and surface-bound antibody tags. The workflow follows a two-step protocol to lyse cells and digest chromatin, making the genome accessible to amplification; the droplets are then subjected to a multiplex PCR to simultaneously amplify the genomic targets and capture antibody tags, labeling them with a droplet barcode relating sequences from the

same cell (**Figure 4.1b**). For genotype, we target recurrently mutated genomic DNA loci in AML with primers containing a unique cell barcode against 50 amplicons spanning 19 genes. The primers and PCR conditions are tuned to enable uniform and quantitative amplification of all targets, since count information is necessary for accurate genotype and immunophenotype characterization. These primers also capture antibody tags from a 23-plex immunophenotyping panel based on those used in clinical minimal residual disease studies[68,69] (**Figure 4.1c**). Sequencing yields a multiomic data set where each cell is represented by two vectors and which can be visualized as a low-dimensional embedding (**Figure 4.1d**).

Peripheral blood mononuclear cells (PBMCs) comprise a diverse and well-understood population easily obtained from a blood draw, and thus provide an excellent sample by which to assess the effectiveness of DAb-seq for mapping hematopoietic immunophenotypes. When applied to PBMCs from a healthy donor, we obtain expected cell subsets across blood compartments, identifying both rare and abundant cells in peripheral blood (**Figure 4.2a,b**). To test single-cell genotyping capability, we also perform DAb-seq on a mixture of three cell lines derived from distinct hematopoietic lineages (Jurkat, Raji, K562) with documented mutations in the targeted genomic regions covered by our single-cell DNA sequencing panel[70]. For all genetic variants, we assign genotype calls to each individual cell: homozygous wildtype, heterozygous alternate, or homozygous alternate. We observe the expected correspondence between single-cell genotypes and phenotypes, as cells of the same genotype segregate within a common immunophenotypic cluster (**Figure 4.2c,d**). Notably, we also find that DAb-seq's genotyping is sufficiently sensitive to differentiate the cells based on zygosity of a given mutation (**Figure 4.2d**). These results show that DAb-seq can simultaneously profile genotype from direct analysis of genomic DNA and immunophenotype from barcoded antibodies.

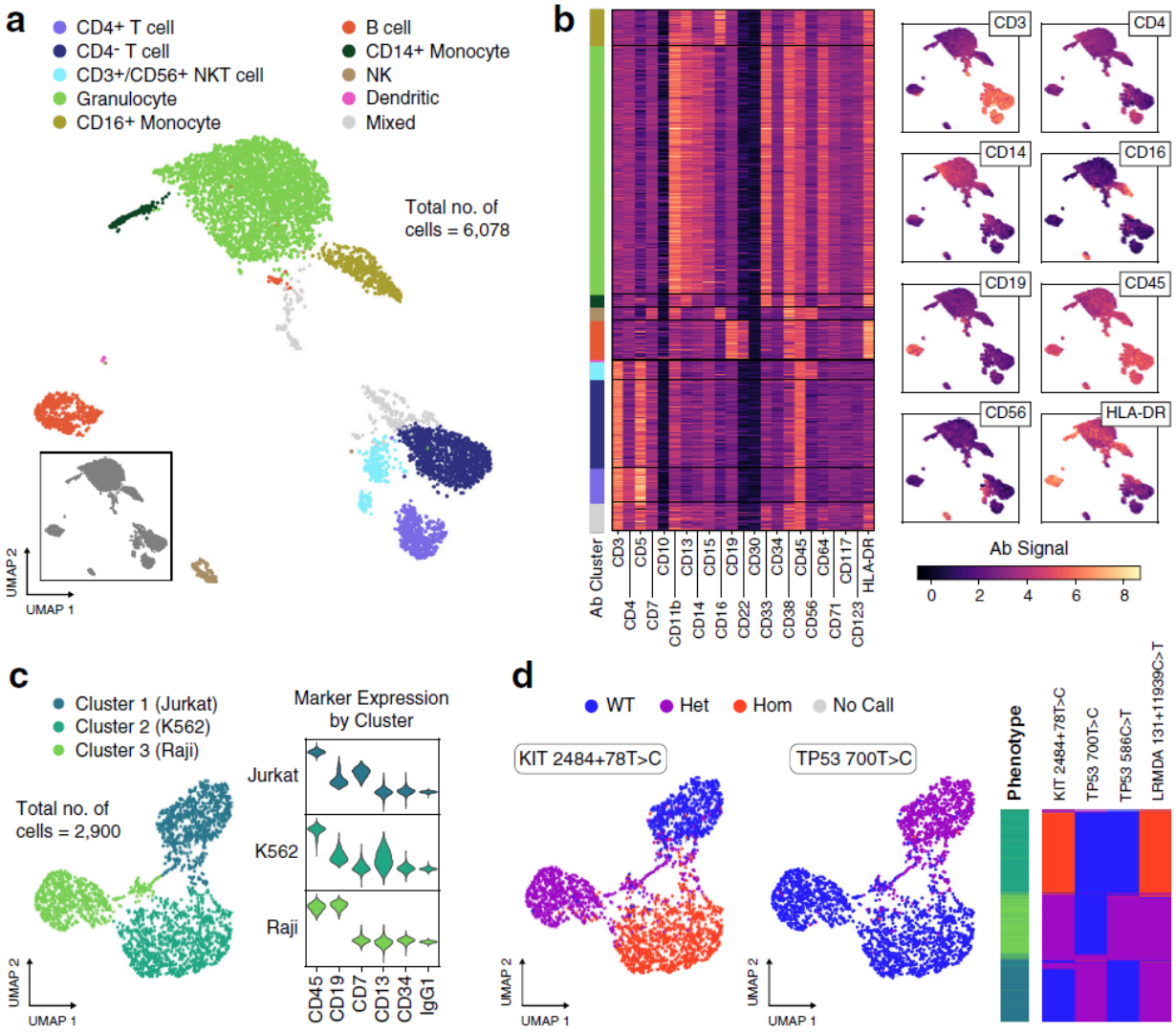

Figure 4.2: DAb-seq enables simultaneous discrimination of single cells by their immunophenotype and genotype.
(a) DAb-seq workflow performed on PBMCs from a healthy donor using a panel of 23 antibodies. Leiden clustering and two-dimensional UMAP embedding of the antibody tag data reveals expected blood compartments. Compartments are annotated based on detected marker expression. **(b)** Heatmap of the corrected log-transformed antibody counts for each cell and antibody. Cells are ordered based on Leiden clusters. Overlay of corrected log-transformed antibody counts with the UMAP embedding highlights compartment-specific expression. **(c)** Correspondence of antibody signal with genomic polymorphisms in DAb-seq experiments tested on a mixture of three cell lines and a panel of six antibodies. Cells cluster by antibody signal as shown in the UMAP embedding. **(d)** Detected single nucleotide polymorphisms in these cells map to the phenotypic cell clusters as shown in the UMAP embedding and a heatmap, where rows correspond to single cells. The first column of the heatmap indicates assigned phenotype cluster, and the remaining columns indicate the genotyping call at the labeled loci.

4.3.2. NPM1-mutated cells persist across therapy timepoints with a static immunophenotype

AML therapies targeted to cell surface proteins require ubiquitous expression of the target marker on the malignant cells. We therefore reason that mutated cells should robustly associate with a common targeted phenotype in patients responsive to this therapy. To investigate this, we perform DAb-seq on 21,952 total cells from bone marrow aspirates of a patient with AML receiving gemtuzumab, a CD33-targeted therapy, across four treatment timepoints (**Figure 4.3a**). This patient received multiple rounds of chemotherapy, including a stem cell transplantation, prior to the first timepoint sampled in this study (**Table 4.1**). In the single-cell DNA genotyping data, we identify a recurrent frameshift mutation in the *NPM1* gene (*NPM1*^{mut}) across relapse, salvage therapy, and progression timepoints. In addition, the *NPM1* mutation is found to always co-occur with a mutation at the *DNMT3A* locus (**Figure 4.3a**). Gemtuzumab targets CD33^{+} cells, which are extinguished at the remission timepoint[71]. To examine the immunophenotypic profile of the *NPM1*^{mut} blast population, we plot single-cell CD33 and CD34 values with *NPM1* mutation status across timepoints (**Figure 4.3b**). The proportion of *NPM1*^{mut} cells in the CD34^{-} and CD34^{+} compartments does not vary extensively across treatments, suggesting the lack of a therapeutic response in the blast immunophenotype. CD33^{+} myeloid cells targeted by the drug are absent at remission.

Table 4.1: AML patient clinical histories.

Gemtuzumab Patient		Pediatric Patient		Gilteritinib Patient	
Date	*Notes*	*Date*	*Notes*	*Date*	*Notes*
Dec-14	Diagnosis	Nov-15	Diagnosis	Oct-18	Diagnosis
	Treatment: cytarabine + idarubicin		Treatment: per study AAML1031 with bortezomib (cytarabine, daunorubicin, etoposide, bortezomib)		Treatment: cytarabine/daunorubicin
Jan-15	End of induction persistent disease	Jan-16	Remission	Nov-18	End of induction therapy (persistent NPM1)
Feb-15	Gilteritinib		Treatment: consolidation per AAML1031 (cytarabine, etoposide, mitoxantrone, bortezomib)	Dec-18	HDAC + glasdigib x2
Mar-15	Persistent disease	Sep-16	Relapse	Jan-19	Persistent NPM1
Apr-15	Treatment: clofarabine, cytarabine			Feb-19	Treatment: azacitidine + venetoclax
May-15	Remission			Mar-19	Recurrence with new FLT3-ITD
Jun-15	HSCT			Apr-19	Treatment: FLAG-Ida
Aug-15	Recurrence			May-19	Recurrence
	Treatment: sorafinib + azacitidine x 6				Treatment: gilteritinib
Jan-16	Treatment: sorafinib maintenance			Jun-19	Prior to HSCT
Mar-19	Relapse				
	Treatment: azacitidine + venetoclax				
Apr-19	End of induction: persistent disease				
	Treatment: azacitidine + venetoclax				
May-19	Progressive disease				
	Treatment: gemtuzumab + gilteritinib				
Jun-19	DLI				
Jun-19	Remission				

Key of acronyms

*HSCT= Hematopoetic stem cell transplant
*DLI = donor lymphocye infusion
*FLAG = fludarabine, cytarabine, G-CSF
*ida = idarubicin

Indicates sample analyzed by DAb-seq

In all timepoints for this patient, our analysis suggests an equivalence between the dominant blast genotype and corresponding phenotype. To further explore this relationship between genotype and phenotype, we visualize the high-dimensional single-cell immunophenotype as a Uniform Manifold Approximation and Projection[72] (UMAP) embedding of the antibody data (**Figure 4.3c**). Cells within single antibody clusters originate from different timepoints, highlighting the stability of normal and malignant immunophenotypes over time.

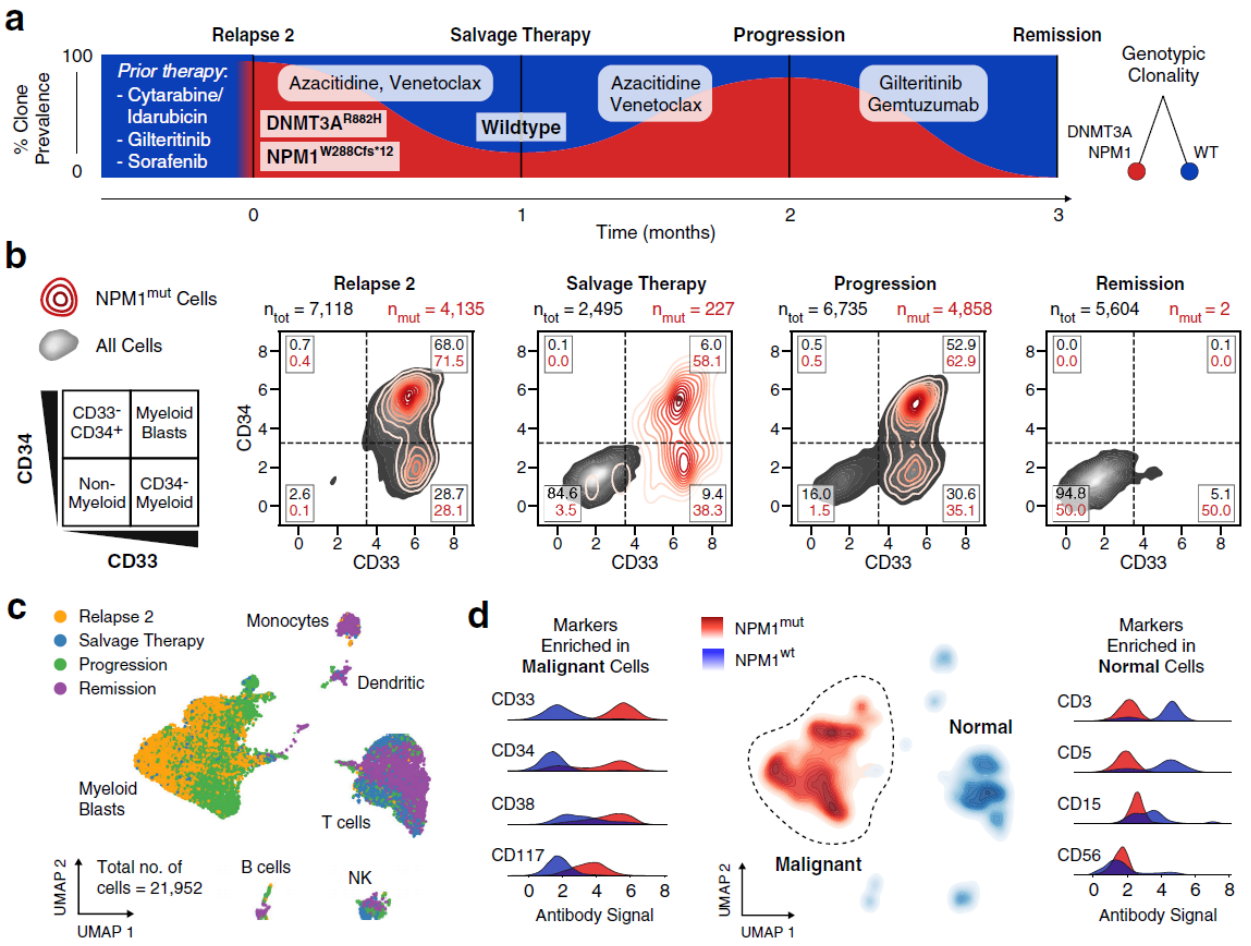

Figure 4.3: AML blasts exhibit a stable genotype and phenotype through treatment.
(a) DAb-seq performed on four bone marrow aspirates of a patient with AML during disease progression as indicated in the fishplot (black lines). The patient received multiple rounds of chemotherapy prior to the experiment (**Table 4.1**). The fraction of blast cells with *NPM1* W288Cfs*12 (NPM1^{mut}) mutation for each sampled time point detected by DAb-seq are shown in red. **(b)** Scatter plots with kernel densities show CD33 and CD34 signal for all cells (grey) and NPM1^{mut} cells (red) for each of the sampled time points. The percentage of normal and mutant cells within each gate are listed. Virtually gating cells highlights a persisting CD33^{+} blast population which is eradicated with gemtuzumab, a CD33-targeted therapy. **(c)** UMAP embedding based on the log-transformed and corrected antibody counts from all cells labeled by timepoint indicates that the high-dimensional immunophenotype of the blasts is stable over the sampled timepoints. **(d)** The genotype of each cell at the NPM1 locus is plotted as a kernel density estimate using the UMAP coordinates from **(c)**. Antibody signals enriched among malignant and normal populations are plotted as kernel densities using all cells and labeled by genotype.

When we overlay *NPM1* genotype on the immunophenotype UMAP space, we find a clear association between a single malignant immunophenotype composed of CD33^{+} cells with *NPM1* mutation status, with variable expression of CD34, CD38, and CD117 in this population

(**Figure 4.3d**). Indeed, this is in agreement with previous observations in flow cytometric studies where blast cells have been found to uniformly express CD33 and variably express CD34, CD38, and CD117[73]. Among the *NPM1*^{wt} cells, we identify classical blood cell markers including CD3 and CD5 (lymphocyte), CD15 (monocyte), and CD56 (natural killer). Taken together, in this patient, DAb-seq confirms elimination of CD33^{+} cells by gemtuzumab treatment and reveals a strong correspondence between genotype and phenotype across timepoints.

4.3.3. Genotypic subclones form overlapping subsets across an immunophenotypic continuum

To investigate whether such tight genotype-phenotype association is a universal feature of AML, we apply DAb-seq to a pediatric patient who underwent induction and consolidation chemotherapy, but ultimately relapsed (**Table 4.1**). We identify two mutually exclusive *KRAS* and *FLT3*-mutated clones at diagnosis and relapse (*KRAS*^{mut}, *FLT3*^{mut}). The *FLT3*^{mut} population, although the minor subclone at diagnosis comprising just 43 of 4,563 cells (0.94%) compared to 1,539 cells (33.7%) for the *KRAS*^{mut} variant, dominates at relapse (6,800 of 7,516 cells, 90.5%) (**Figure 4.4a**). Immunophenotypically, we also identify a third subset comprising *KRAS*^{WT}/*FLT3*^{WT} cells expressing a blast-like CD33^{+}CD38^{+} immunophenotype with no identifiable DNA mutations in the targeted loci. When we group cells from all timepoints by genotype, pathogenic blasts display variable patterns in immunophenotype, with no clear mapping between the two (**Figure 4.4b**).

In the absence of an obvious genotype-phenotype mapping for this patient, we sought to investigate the underlying relationship between these domains. Using UMAP, we project the antibody data into two dimensions, coloring the points according to genotype (**Figure 4.4c**). We observe a single immunophenotypic compartment with incomplete separation between genotypes. To estimate antibody profile expression within the blast compartment continuum, we identify the dominant gradient in the phenotypic space, ordering all points along the gradient. We then calculate the local average antibody and genotypic composition for neighboring cells

(**Figure 4.4c,d**) (also see section 4.5: Methods). As expected, many markers are anticorrelated (CD11b, CD33, CD56) or correlated (CD15) with the principal immunophenotypic gradient. Less trivially, genotypic compositions vary along the gradient, with *KRAS*^{mut} clone frequencies anticorrelated and *FLT3*^{mut} correlated (**Figure 4.4d**). Nevertheless, genotype composition never completely separates into individual clonal populations, making it impossible to define distinct genotype-phenotype clusters; consequently, technologies profiling one modality, such as genotyping or immunophenotyping, cannot adequately capture the heterogeneity inherent to this case of AML.

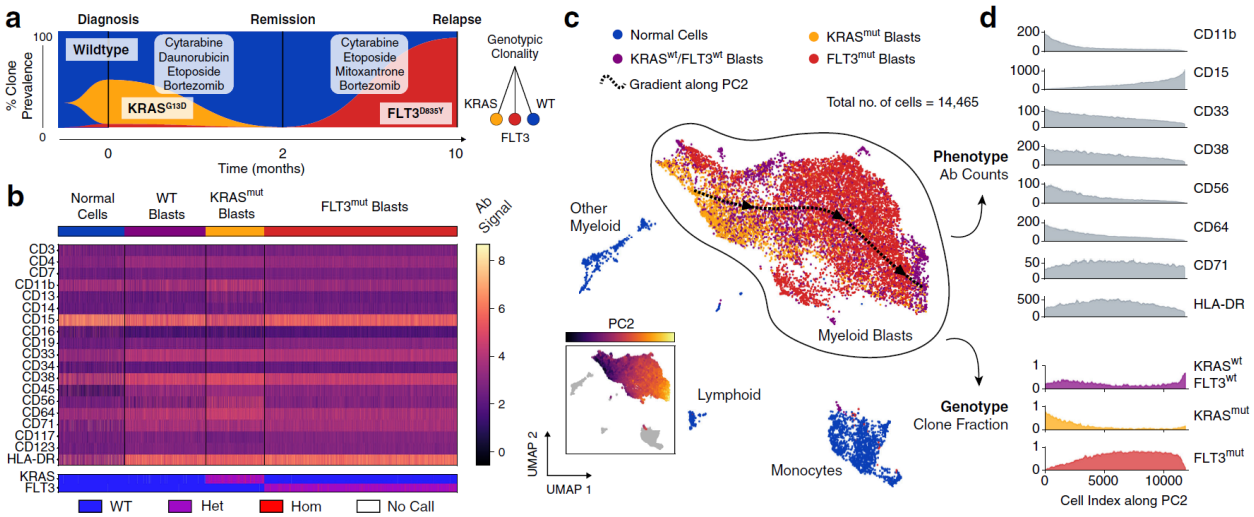

Figure 4.4: Distinct genetic subclones form an overlapping immunophenotypic continuum in a case of pediatric AML.
(a) Three timepoints sampled with DAb-seq during treatment comprise a mixture of independent clones (KRAS G13D heterozygous blasts, yellow; FLT3 D835Y blasts, red). The wildtype compartment contains additional cells with a blast-like immunophenotype lacking detectable mutations. **(b)** Heatmap of log-transformed corrected antibody counts and genotyping calls for the KRAS and FLT3 loci for each cell across all timepoints. The heatmap is grouped by genotype. Cells with wildtype genotype but blast-like immunophenotype are labeled separately. **(c)** UMAP embedding of all cells from all time points based on log-transformed corrected antibody counts. Color indicates mutation status as in a. The blast compartment is overlaid with a spline approximating the gradient of the 2nd principal component of the antibody count matrix (shown in inlet figure) and indicates a gradual change in immunophenotype. **(d)** Moving average expression of antibodies and fraction of mutated cells sorted by the 2nd principal component of the antibody count matrix. The overlapping phenotypic continuum between the genetically distinct blast clones is apparent.

4.3.4. FLT3 inhibitor therapy induces erythroid differentiation in a case of AML

Our first two cases feature either a strong genotype-phenotype correlation (Patient 1) or mixed genotyping comprising a single immunophenotype (Patient 2). Thus, for our final case, we analyze a patient treated with gilteritinib, a FLT3 inhibitor therapy reported to promote *in vivo* differentiation of myeloid blasts. This treatment is thought to disperse distinct genotypes into multiple immunophenotypes, although the terminal lineage of the cells remains poorly understood[74–76]. Accordingly, we hypothesize DAb-seq should allow tracking of immunophenotypic dispersal and confirmation of their terminal hematopoietic lineage. We analyze 18,287 cells across treatment timepoints, beginning at diagnosis, discovering a subclone with co-mutated *DNMT3A* and *NPM1* (**Figure 4.5a**; **Table 4.1**). Following cytarabine/daunorubicin induction therapy, a fraction of *DNMT3A*^{mut} cells remain at remission. At relapse and after treatment with the FLT3 inhibitor gilteritinib ("FLT3 Inhibitor"), most cells contain a 24-bp *FLT3* internal tandem duplication (ITD), in addition to the initial *DNMT3A* and *NPM1* mutations. The genotypic structure inferred from the single-cell data indicates a linear, branching hierarchy of sequentially acquired mutations in response to therapy. To explore the immunophenotypic features of this patient's disease, we integrate cells from all timepoints and construct a UMAP representation using the antibody data (**Figure 4.5b**). We cluster this data using the Leiden method for cluster detection, an improved algorithm over Louvain modularity[77,78], and manually annotate with phenotypic labels corresponding to hematopoietic lineage from the antibody data (**Figure 4.5c**). We identify three blast populations expressing high levels of CD33 and CD38, a monocytic population expressing CD15 and CD16, and erythroid and lymphoid clusters with elevated CD71 and CD3. As expected, samples across treatment timepoints comprise a mixture of immunophenotypically normal and blast-like cells.

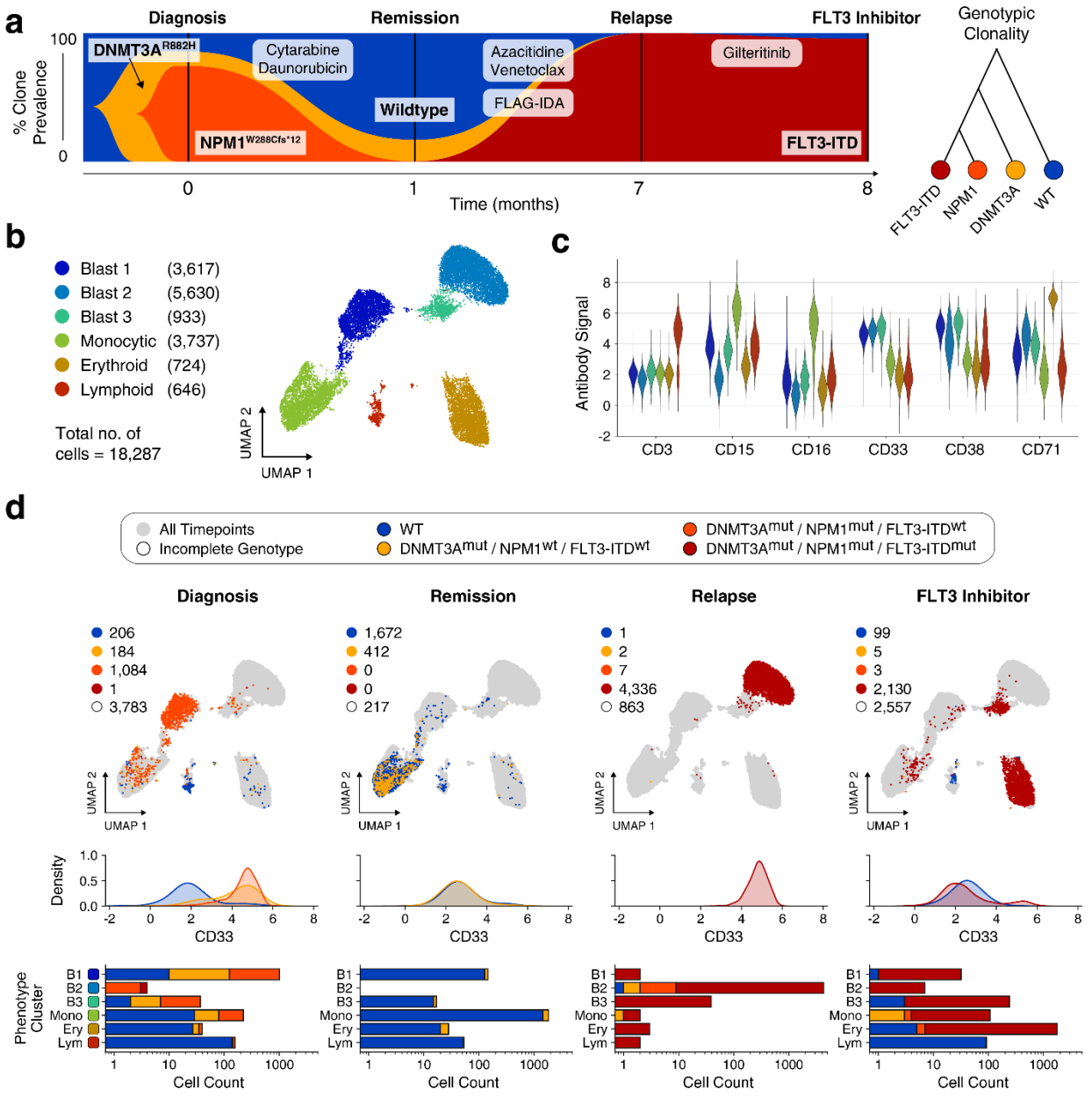

Figure 4.5: Decoupling of blast phenotype and genotype in response to FLT3 inhibitor therapy. **(a)** Fishplot showing observed fraction of cells with distinct genetic mutations for each sampled time point. The co-occurrence of the three mutations in the single-cell data is consistent with a linear model of mutation accumulation. **(b)** UMAP embedding of all cells based on measured antibody signal. The cells segregate into six distinct phenotypic clusters with multiple blast compartments. **(c)** Average expression of each cell cluster for a selection of markers. **(d)** Top row: Same UMAP embedding as in b given as grey outline. For each sampled time point, observed cells are plotted and colored according to the detected genotype. Blasts distribute among multiple phenotypic compartments in the final time point following FLT3 inhibitor treatment. Middle row: Kernel density plot of the CD33 antibody signal resolved by time point and genotype. Cells from genotypic compartments with less than 10 cells per time point are not plotted. Bottom row: Bar chart depicting genotypic composition of each phenotypic cluster in b resolved by time point.

Hypothesizing that different therapies should yield different genotype-phenotype coupling patterns, we sought to characterize how mutated and normal cells distribute across immunophenotypic clusters. For each timepoint, we thus label cells in UMAP space according to DNA genotype and generate density distributions of CD33 signal, a pan-myeloid marker (**Figure 4.5d**). We also evaluate counts of phenotype cluster membership in each timepoint, subdivided by DNA genotype. At diagnosis, cells mutated at both the *DNMT3A* and *NPM1* locus reside primarily in the Blast 1 cluster (81.8% of *DNMT3A*^{mut}/*NPM1*^{mut} cells) and express high levels of CD33. A secondary clone mutated exclusively at the *DNMT3A* locus exhibits comparable CD33 expression and resides mainly in the Blast 1 and monocytic clusters (62.5% and 27.7% of *DNMT3A*^{mut} cells, respectively). At remission, the same *DNMT3A*^{mut} clone is identified but with decreased CD33 expression and a primarily monocytic immunophenotype (92.7% of *DNMT3A*^{mut} cells) co-localizes with cells of normal genotype, consistent with clonal hematopoiesis of a pre-leukemic clone[79,80]. A newly acquired *FLT3*-ITD clone emerges in high numbers at relapse (99.8% of genotyped cells), coinciding with a phenotypic shift of cells to the CD33^{+} Blast 2 cluster. Following FLT3 inhibitor treatment, the same *FLT3*-ITD clone persists but exhibits a transformed immunophenotype, as evidenced by membership of the *FLT3* clone in multiple immunophenotypic clusters. The new *FLT3*-ITD immunophenotype is primarily erythroid (82.2% of *FLT3*-ITD cells), with minor fractions in the Blast 3 and monocytic compartments (11.1% and 4.84% of *FLT3*-ITD cells, respectively). Furthermore, the *FLT3*-ITD clone at relapse lacks uniform CD33 expression, indicating that this clone is no longer restricted to the myeloid compartment. Taken together, these findings support the model of terminal erythroid differentiation of blasts in a case of leukemia treated with gilteritinib. In agreement with a recent study[76], proteogenomic analysis by DAb-seq challenges a prior report of gilteritinib-induced terminal differentiation towards a myeloid fate[75]. DAb-seq elucidates the rich and complex dynamics of this process and illustrates how distinct DNA genotypes can fractionate into multiple phenotypic identities in response to treatment.

4.4. Discussion

Through its ability to jointly profile DNA and immunophenotype, DAb-seq captures the complexity of proteogenomic states underlying AML. Analysis of multiple patients over timepoints and treatments demonstrates the plasticity of the disease and the complex and unpredictable way it progresses in different contexts. In a patient with extensive clinical history including multiple rounds of chemotherapy, we found a robust relationship between mutant *NPM1* cells and a malignant phenotype; this suggested that a single CD33-targeted therapy would eradicate the blast population, as indeed it did. By contrast, in a separate case of pediatric AML, we observed that genetically distinct populations shared overlapping immunophenotype, demonstrating that this domain alone is insufficient for characterizing how cells are genetically programmed and may, consequently, respond to treatment. In the final case study, we observed the opposite scenario, in which treatment by gilteritinib induced mutationally similar cells to disperse into different myeloid compartments, highlighting the challenge of targeting these malignant cells for eradication. Our results thus demonstrate that genotype or immunophenotype alone is insufficient to predict the evolution of proteogenomic states in AML.

DAb-seq employs targeted primers to amplify specific genomic regions and panels of antibodies. While both readouts enable massive multiplexing of queried targets, practical and economic constraints necessitate *a priori* knowledge of which loci and epitopes to profile. As such, the strength of DAb-seq is not unbiased feature discovery, as with scRNA-seq, but rather sensitive and precision analysis of actionable information. Furthermore, as with all targeted methods of DNA genotyping, DAb-seq cannot exclude the possibility that disease-relevant mutations occur beyond the sequenced loci or in immunophenotypic markers not included in the panels. In the case of pediatric AML, it is therefore impossible for us to conclude if the *FLT3*^{wt}*/KRAS*^{wt} blast population is driven by epigenetic changes or unmapped genomic aberrations. Nevertheless, the sensitivity of DAb-seq, and its low genotyping drop-out, allows

identification of co-occurring mutations, including heterozygous mutations that are notoriously difficult for RNA-based approaches. Moreover, DAb-seq firmly places genomic mutations in understood phenotypic contexts, which is vital for understanding how they program the disease and, ultimately, treatments select for them.

In the era of personalized medicine, treatment decisions are increasingly based on DNA mutation status, such as targeted EGFR inhibitors or protein expression like HER2 or PD-L1 status. To fully leverage the capabilities of modern profiling techniques, however, information across all available domains must be integrated to optimize the therapeutic strategy for a given patient. Indeed, our findings underscore the importance of utilizing both genotype and immunophenotype to fully characterize disease and assess efficacy of treatment. For example, CAR T-cell therapy derives specificity from protein expression, yet would fail to elicit a complete response if pathogenic genotypes were distributed across multiple phenotypic clusters. Such a scenario would require joint single-cell profiling as in DAb-seq to unravel. As multiomic single-cell technologies like DAb-seq become available, it will be feasible to use comprehensive precision analysis to deconvolute the subtlety of each patient's cancer and thereby select the best treatment regimen.

4.5. Methods

4.5.1. Conjugation of antibodies to oligonucleotide barcodes

Monoclonal antibodies were conjugated to azide-modified oligonucleotides using a copper-free click chemistry reaction as described previously[81]. Monoclonal antibodies were resuspended to 100 µg in 100 µL PBS. Antibodies were incubated with DBCO-PEG5-NHS Ester linker (Click Chemistry Tools, cat. no. A102P) at a 4:1 molar ratio linker:antibody for 2 h at room temperature. Following incubation, the antibody-linker solution was washed once in a 50 kDa cellulose spin filter (Millipore Sigma, cat. no. UFC505024). DNA oligonucleotides with a 5' azide modification (Integrated DNA Technologies) were reconstituted in water and added to the

washed antibodies at a 2.5:1 molar ratio oligonucleotide:antibody. Following a 16 h incubation, the conjugated antibodies were washed three times in a 50 kDa filter to remove unreacted oligonucleotides. All antibody conjugates were run on a Bioanalyzer Protein 230 electrophoresis chip (Agilent Technologies, cat. no. 5067-1517) to verify successful conjugation.

4.5.2. Cell culture and PBMC processing for control experiments

The following three cell lines were used in the initial control experiment: Raji (ATCC, CCL-86), Jurkat (ATCC, TIB-152), K562 (ATCC, CCL-243). Cells were cultured under the supplier's recommended conditions. PBMCs from a single healthy donor were sourced commercially (iXCells Biotechnologies, cat. no. 10HU-003) and stored at -80°C until use. Prior to staining, the cultured cell lines and PBMCs were washed once in PBS with 5% fetal bovine serum (FBS) (Thermo Fisher, cat. no. 10082147). For the control experiment, the three cell lines were combined at an equal ratio.

4.5.3. Collection of patient samples

Patients included in this study were treated at the University of California, San Francisco (UCSF), and peripheral blood or bone marrow was stored in the UCSF tumor bank. Samples were processed immediately after collection to isolate mononuclear cells. Sample collection was in accordance with the Declaration of Helsinki under institutional review board-approved tissue banking protocols. Written informed consent was obtained from all patients.

4.5.4. Thawing patient samples

A protocol was optimized to maximize recovery of viable cells from patient samples. Cryovials containing patient tissue (peripheral blood or bone marrow aspirate) were warmed by hand and carefully transferred dropwise to a 50 mL tube containing 40 mL of cold DMEM media (Thermo Fisher, cat. no. 11995040) with 20% FBS and 2 mM EDTA. The tube was centrifuged at 700 rpm at 4°C for 7 min with no brake. The supernatant was discarded, and the cells were

resuspended in 10 mL of warmed RPMI-1640 media (Thermo Fisher, cat. no. A1049101) with 10% FBS. The solution was strained through a 70 µm cell strainer (Corning, cat. no. 431751) to remove any large cell aggregates and the tube was centrifuged a second time at 700 rpm at 4°C for 5 min with low brake. The supernatant was discarded, and the cells were resuspended in PBS with 5% FBS for staining.

4.5.5. Cell staining using oligonucleotide-conjugated antibodies

For each sample, 2 million cells were added to a 5 mL DNA LoBind tube (Eppendorf, cat. no. 0030108310), centrifuged at 400 x g for 4 min, and resuspended in 180 µL PBS with 5% FBS. Cells were blocked for 10 min on ice following addition of 10 µL Fc blocking solution (BioLegend, cat. no. 422301), 4 µL of a 1% dextran sulfate solution (Research Products International, cat. no. D20020), and 4 µL of 10 mg/mL salmon sperm DNA (Invitrogen, cat. no. 15632011). Cells were stained for 30 min on ice with 0.5 µg of each conjugated antibody. After incubation, five rounds of washing were performed to remove excess antibody. For each wash, 5 mL PBS with 5% FBS was added to the tube and centrifuged at 400 x g for 4 min. Stained cells were resuspended in Mission Bio cell buffer at a final concentration of 3 M/mL prior to microfluidic encapsulation.

4.5.6. Microfluidic single-cell DNA genotyping and antibody capture

A commercial single-cell DNA genotyping platform (Mission Bio, Tapestri) was used to perform microfluidic encapsulation, lysis, and barcoding according to the manufacturer's protocol for the acute myeloid leukemia V1 panel. Where noted, modifications were made to enable co-capture of oligonucleotide-labeled antibodies. Stained cells were loaded into a microfluidic cartridge and co-encapsulated into droplets with a lysis buffer containing protease and mild detergent. Droplets were incubated in a thermal cycler for 1 h at 50°C to digest all cellular proteins, followed by 10 min at 80°C to heat-inactivate the protease. To enable antibody capture during the barcoding stage, the antibody tags were designed with 3' complementarity to

one of the *RUNX1* gene forward primers and the corresponding reverse primer was omitted from the reverse primer pool. Lysed cells in droplets were transferred to the barcoding module of the microfluidic cartridge in addition to polymerase mix, the modified reverse primer pool, barcoded hydrogel beads, and oil for droplet generation. The droplets were placed under a UV lamp (Analytik Jena, Blak-Ray XX15L) for 8 min to cleave the single-stranded PCR primers containing unique cell barcodes from the hydrogel beads. To amplify DNA targets and capture antibody tags, droplets were thermal cycled using the following program: 95°C for 10 m; 20 cycles of (95°C for 30 s, 72°C for 10 s, 61°C for 4 min, 72°C for 30 s); 72°C for 2 min; 4°C hold.

4.5.7. Single-cell DNA amplicon and antibody tag sequencing library preparation

Recovery and cleanup of single-cell libraries proceeded according to the Mission Bio V1 protocol with additional modifications for antibody library preparation. The 8 PCR tubes containing barcoded droplets were pooled as pairs and treated with Mission Bio Extraction Agent. Water was added to each tube and the aqueous fraction transferred to a new 1.5 mL DNA LoBind tube. Ampure XP beads (Beckman Coulter, cat. no. A63881) were added at a 0.75X volume ratio beads:PCR product for size selection. The supernatant from the size selection step, containing library fragments shorter than ~200 bp, was retained and used for antibody library preparation, while the remaining beads with bound DNA panel library fragments were washed twice with 80% EtOH and eluted in 30 µL water. A biotinylated capture oligonucleotide (/5Biosg/GGCTTGTTGTGATTCGACGA/3C6/, Integrated DNA Technologies) complementary to the 5’ end of the antibody tags was added to the retained supernatant to a final concentration of 0.6 µM. The supernatant-probe solution was heated to 95°C for 5 min to denature the PCR product, then snap-cooled on ice for probe hybridization. 10 µL of streptavidin beads (Thermo Fisher, cat. no. 65001) were washed according to the manufacturer’s protocol and added to each tube of PCR product. Following a 15 min incubation at room temperature, the beads were isolated by magnetic separation, washed two times in

PBS, and resuspended in 30 µL water. PCR was performed on the purified DNA panel and antibody tags to produce sequencing libraries. For each tube of purified DNA panel, 50 µL reactions were prepared containing 4 ng of barcoded product in 15 µL water, 25 µL Mission Bio Library Mix, and 5 µL each of custom P5 and Nextera P7 primers (N7XX), both at 4 µM stock concentration. The reactions were thermal cycled using the following program: 95°C for 3 min; 10 cycles of (98°C for 20 s, 62°C for 20 s, 72°C for 45 s); 72°C for 2 min; 4°C hold. For each tube of purified antibody tags, identical reactions were prepared, instead using 15 µL bead-bound template, 5 µL antibody tag-specific P7 primer at 4 µM, and 20 cycles of amplification. Following amplification, both the DNA panel and antibody tag libraries were cleaned with 0.7X Ampure XP beads and eluted in 12 µL water.

4.5.8. Next-generation sequencing

All DNA panel and antibody tag libraries were run on a Bioanalyzer High Sensitivity DNA electrophoresis chip (Agilent Technologies, cat. no. 5067-4626) to verify complete removal of primer-dimer products. Libraries were quantified by fluorometer (Qubit 3.0, Invitrogen) and sequenced on Illumina next-generation sequencing platforms with a 20% spike-in of PhiX control DNA (Illumina, cat. no. FC-110-3001). All sequencing runs used a dual-index configuration and a custom Read 1 primer (5’ GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTAG 3’, Integrated DNA Technologies). The 3-cell control sample was sequenced on an Illumina MiSeq using a v2 300-cycle kit in 2x150 bp paired-end mode (Illumina, cat. no. MS-102-2002). For the patient samples, DNA panel and antibody tag libraries were sequenced separately to maximize cost-effectiveness. DNA panels were sequenced with an Illumina NovaSeq 6000 SP 300-cycle Kit (Illumina, cat. no. 20027465) in 2 x 150 bp paired-end mode. Antibody tag libraries were sequenced with an Illumina NextSeq 550 75-cycle High Output Kit (Illumina, cat. no. 20024906) in paired-end mode, using 38 cycles for Read 1 and 39 cycles for Read 2.

4.5.9. Bioinformatic pipeline for single-cell DNA genotyping and antibody tag counting

Sequencing data was processed using a custom pipeline available on GitHub (see Code Availability). For all reads, combinatorial cell barcodes were parsed from Read 1 using cutadapt (v2.4) and matched to a barcode whitelist. Barcode sequences within a Hamming distance of 1 from a whitelist barcode were corrected.

For the DNA genotyping libraries, reads with valid barcodes were trimmed with cutadapt to remove 5' and 3' adapter sequences and demultiplexed into single-cell FASTQ files using the script "demuxbyname" from the BBMap package (v.38.57). Valid cell barcodes were selected using the inflection point of the cell rank plot in addition to the requirement that 60% of DNA intervals were covered by a minimum of 8 reads. FASTQ files for valid cells were aligned to the hg19 build of the human genome reference using bowtie2 (v2.3.4.1). The single-cell alignments in BAM format were filtered (properly mapped, mapping quality > 2, primary alignment), sorted, and indexed with samtools (v1.8). GVCF files were produced for all cells using HaplotypeCaller from the GATK suite (v.4.1.3.0). Joint genotyping was performed on all genomic intervals in parallel (excluding primer regions) using GATK GenotypeGVCFs. For longitudinal patient samples, cells from all timepoints were joint genotyped as a multi-sample cohort. Genotyped intervals from all cells were combined into a single variant call format (VCF) file, and multiallelic records were split and left-aligned using bcftools (v1.9). Variants were annotated with ClinVar metadata (v.20190805) and SnpEff functional impact predictions (v4.3t). Variant records for all cells were exported to HDF5 format using a condensed representation of the genotyping calls (0: wildtype; 1: heterozygous alternate; 2: homozygous alternate; 3: no call).

The antibody tag libraries were processed identically for cell barcode demultiplexing. For reads with valid cell barcodes, 8 bp antibody barcodes and 10 bp unique molecular identifiers (UMIs) were extracted from Read 2 using cutadapt with the requirement that all UMI bases had a minimum quality score of 20. Antibody barcode sequences within a Hamming distance of 1 from known antibody barcodes were corrected. UMI sequences were grouped by cell and

antibody and counted using the UMI-tools package (v.0.5.3, “adjacency” method). UMI counts of antibodies for each cell barcode were exported in tabular format for further analysis.

4.5.10. Cell and genotype filtering

Cell barcodes were additionally filtered according to antibody counts. Valid barcode groups were required to have a minimum of 100 antibody UMIs by the adjacency counting method and a maximum IgG1 count no greater than five times the median IgG1 count of the associated DAb-seq experiment. For each valid cell barcode, all variants were filtered according to the quality and sequence depth reported by GATK. Genotyping calls were required to have a minimum quality of 30 and total depth of 10; variant entries below these thresholds were marked as “no call” and excluded from analyses.

4.5.11. Antibody-based embedding and clustering

To correct for technical effects in the raw antibody counts and batch variability between experiments from the same patient but different time points, a linear regression over all cells from the same patient was performed. This approach is similar to those employed in single-cell RNA-seq normalization procedures[82], which attempt to reduce technical and biological noise in expression data by treating variables (sequencing depth, cell cycle, etc...) as regressors in a linear model. For DAb-seq antibody data, to all entries c_{ij} of the UMI corrected antibody count matrix c, where *i* is the cell index and *j* the antibody index, one pseudocount was added and the matrix was log-transformed. A matrix of quality metrics q with cells as rows and four columns (total antibody reads, total antibody counts after UMI correction, IgG1 count and total amplicon reads) was log-transformed, column-wise normalized, and mean-centered. A singular value decomposition was performed on the transformed matrix q and the left-singular vectors retained as design matrix. Each column vector c_j was then regressed with either the first three, two, or one left-singular vectors, for patient samples, PMBC or cell lines respectively as regressors. The vector of residuals u_j is then the corrected antibody signal of antibody *j* (**Figure 4.6**).

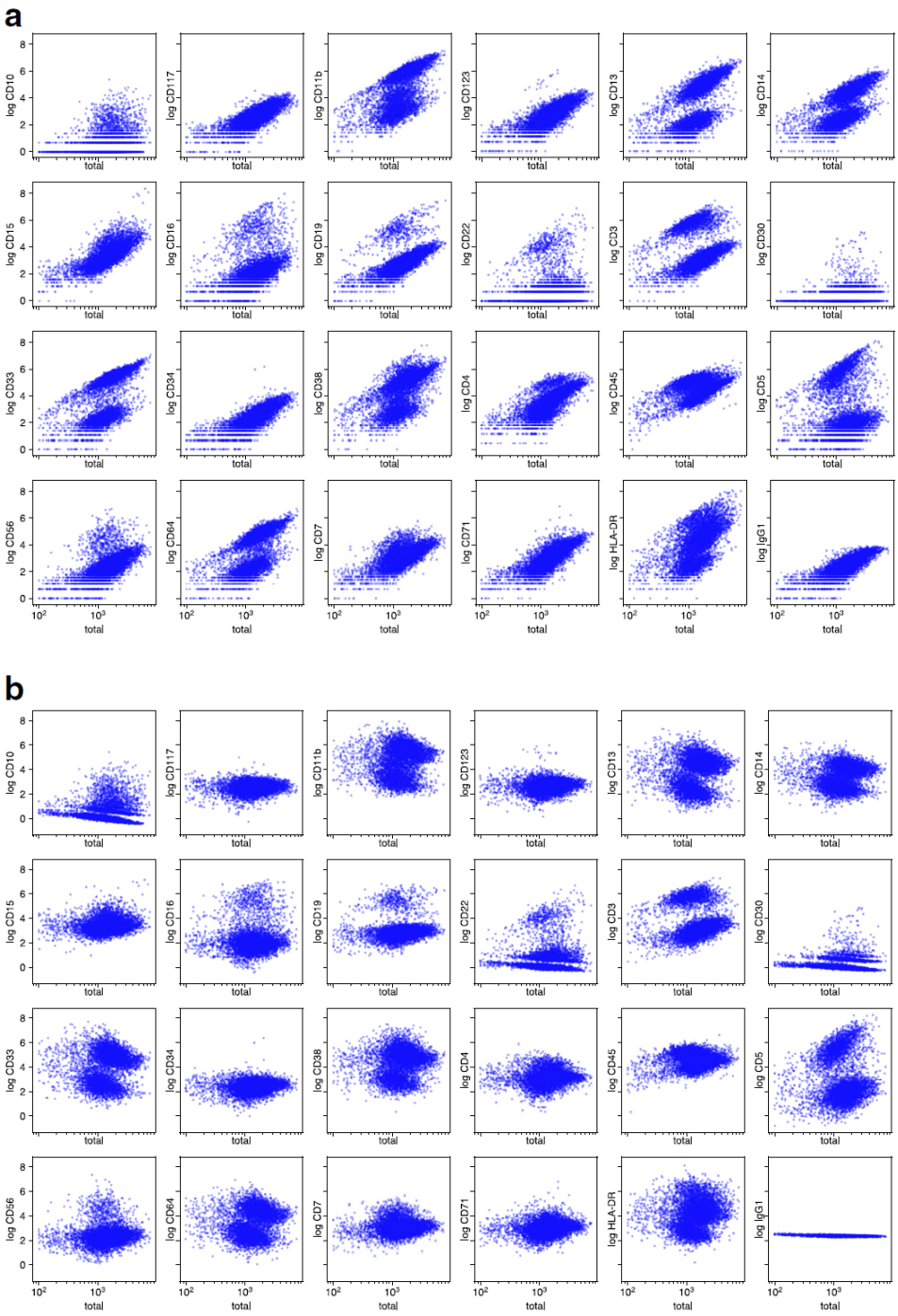

Figure 4.6: Antibody count bias correction by linear regression.
(a) Raw UMI counts for each antibody and cell are plotted versus total antibody count from the same cell. A clear correlation between the two is visible. A similar slope is visible for the isotype control (bottom row, rightmost column), suggesting technical bias. **(b)** Same plots as in (a) after correcting for global droplet performance by linear regression. Correlation with total antibody counts is reduced. In both corrected and uncorrected plots for many of the markers, two clusters of cells are prominent, representing low- and high-expressing cells.

A UMAP embedding in two dimensions of the corrected antibody signal was done in Python 2.7 using the umap-learn[72] (v0.3.10) and scanpy[83] (v.1.4.4.post1) packages, with the minimum distance parameter set to 0.1 for the pediatric patient and 0.2 for all other samples and default parameters otherwise. To construct the underlying nearest neighbor graph from the corrected antibody count matrix, 15 or 16 nearest neighbors based on the first 16 to all principal components were used. The scanpy implementation of the Leiden algorithm[77] with resolution set to 0.1 for the three cell line experiment and 1 otherwise was used to assign cells to phenotypic compartments. A comparison of single-cell UMAP plots derived from raw and linear regression-corrected antibody counts is shown in **Figure 4.7**. The corrected antibody counts produce less dispersed, discrete cell clusters due to a reduction in noise from amplification bias and nonspecific antibody binding.

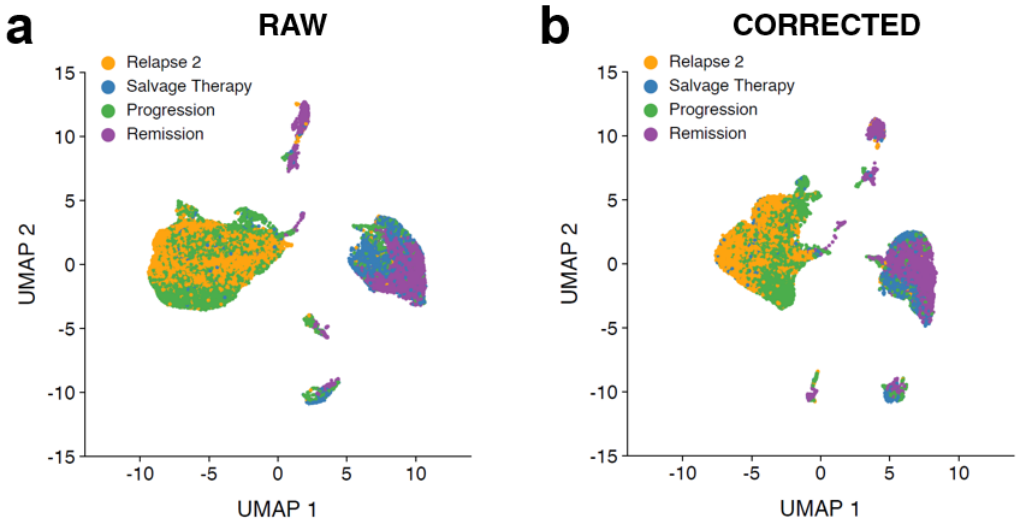

Figure 4.7: Single-cell UMAP plots derived from raw and corrected antibody counts for the patient treated with CD33-targeted therapy.
(a) UMAP plot of raw single-cell antibody counts, using only the UMI-corrected values. (b) UMAP plot of single-cell antibody counts, corrected by linear regression.

For the gradient analysis of the pediatric patient with AML (**Figure 4.4**), only cells belonging to Leiden communities with blast phenotype were retained and the singular value

decomposition of the remaining rows of u was calculated. Cells were then ordered by their value of the second left-singular vector. Antibody counts and genotype fractions along the gradient were averaged with a moving window of 200 cells. Similarly, the average position of the cells in the two-dimensional UMAP embedding was estimated by smoothing x and y coordinates with a moving window of the same length. A 3rd-order spline was placed through the smoothed cell position to indicate the orientation of the gradient in the UMAP embedding.

References

1. Lan, F., Demaree, B., Ahmed, N. & Abate, A. R. Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. *Nat. Biotechnol.* **35**, 640–646 (2017).
2. Demaree, B., Weisgerber, D., Lan, F. & Abate, A. R. An Ultrahigh-throughput Microfluidic Platform for Single-cell Genome Sequencing. *J. Vis. Exp.* e57598–e57598 (2018). doi:10.3791/57598
3. Demaree, B., Weisgerber, D., Dolatmoradi, A., Hatori, M. & Abate, A. R. Direct quantification of EGFR variant allele frequency in cell-free DNA using a microfluidic-free digital droplet PCR assay. in *Methods in Cell Biology* **148**, 119–131 (Academic Press Inc., 2018).
4. Demaree, B. *et al.* Joint profiling of proteins and DNA in single cells reveals extensive proteogenomic decoupling in leukemia. *bioRxiv* 2020.02.26.967133 (2020). doi:10.1101/2020.02.26.967133
5. Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell. *Science (80-.).* **338**, 1622–1626 (2012).
6. McConnell, M. J. *et al.* Mosaic copy number variation in human neurons. *Science (80-.).* **342**, 632–637 (2013).
7. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90-U119 (2011).
8. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
9. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
10. Rotem, A. *et al.* High-throughput single-cell labeling (Hi-SCL) for RNA-Seq using drop-

based microfluidics. *PLoS One* **10**, 1–14 (2015).

11. DNA Sequencing Costs: Data | NHGRI. Available at: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data. (Accessed: 23rd March 2020)
12. Del Monte, U. Does the cell number 109 still really fit one gram of tumor tissue? *Cell Cycle* **8**, 505–506 (2009).
13. Maranger, R. & Bird, D. Viral abundance in aquatic systems:a comparison between marine and fresh waters. *Mar. Ecol. Prog. Ser.* **121**, 217–226 (1995).
14. Zhang, H. & Liu, K.-K. Optical tweezers for single cells. *J. R. Soc. Interface* **5**, 671–690 (2008).
15. Rinke, C. *et al.* Obtaining genomes from uncultivated environmental microorganisms using FACS–based single-cell genomics. *Nat. Protoc.* **9**, 1038–1048 (2014).
16. Gawad, C., Koh, W. & Quake, S. R. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proc. Natl. Acad. Sci.* **111**, 17947–17952 (2014).
17. Leung, K. *et al.* Robust high-performance nanoliter-volume single-cell multiple displacement amplification on planar substrates. *Proc. Natl. Acad. Sci.* **113**, 8484–8489 (2016).
18. Tamminen, M. V. & Virta, M. P. J. Single gene-based distinction of individual microbial genomes from a mixed population of microbial cells. *Front. Microbiol.* **6**, 195 (2015).
19. Podar, M. *et al.* Targeted access to the genomes of low-abundance organisms in complex microbial communities. *Appl. Environ. Microbiol.* **73**, 3205–3214 (2007).
20. Xu, L., Brito, I. L., Alm, E. J. & Blainey, P. C. Virtual microfluidics for digital quantification and single-cell sequencing. *Nat. Methods* **13**, 759–762 (2016).
21. Lan, F., Haliburton, J. R., Yuan, A. & Abate, A. R. Droplet barcoding for massively parallel single-molecule deep sequencing. *Nat. Commun.* **7**, 11784 (2016).
22. Rotem, A. *et al.* Single-cell ChIP-seq reveals cell subpopulations defined by chromatin

state. *Nat. Biotechnol.* **33**, 1165–1172 (2015).

23. Novak, R. *et al.* Single-cell multiplex gene detection and sequencing with microfluidically generated agarose emulsions. *Angew. Chemie - Int. Ed.* **50**, 390–395 (2011).
24. Garstecki, P. *et al.* Formation of monodisperse bubbles in a microfluidic flow-focusing device. *Appl. Phys. Lett.* **85**, 2649–2651 (2004).
25. Amini, S. *et al.* Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* **46**, 1343–1349 (2014).
26. De Bourcy, C. F. A. *et al.* A quantitative comparison of single-cell whole genome amplification methods. *PLoS One* **9**, (2014).
27. Gole, J. *et al.* Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. *Nat. Biotechnol.* **31**, 1126–1132 (2013).
28. Nathan, C. & Cars, O. Antibiotic resistance - Problems, progress, and prospects. *N. Engl. J. Med.* **371**, 1761–1763 (2014).
29. Yildiz, F. H. Processes controlling the transmission of bacterial pathogens in the environment. *Research in Microbiology* **158**, 195–202 (2007).
30. Jiang, S. C. & Paul, J. H. Gene transfer by transduction in the marine environment. *Appl. Environ. Microbiol.* **64**, 2780–2787 (1998).
31. Ochman, H. & Moran, N. A. Genes lost and genes found: Evolution of bacterial pathogenesis and symbiosis. *Science* **292**, 1096–1098 (2001).
32. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
33. Afshinnekoo, E. *et al.* Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics. *Cell Syst.* **1**, 72–87 (2015).
34. Iverson, V. *et al.* Untangling genomes from metagenomes: Revealing an uncultured class of marine euryarchaeota. *Science (80-.).* **335**, 587–590 (2012).
35. Ni, X. *et al.* Reproducible copy number variation patterns among single circulating tumor

cells of lung cancer patients. *Proc. Natl. Acad. Sci.* **110**, 21083–21088 (2013).

36. Liu, B. & Pop, M. ARDB-Antibiotic Resistance Genes Database. *Nucleic Acids Res.* **37**, 443–447 (2008).

37. Chen, L., Xiong, Z., Sun, L., Yang, J. & Jin, Q. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Res.* **40**, D641-5 (2012).

38. da Cunha Santos, G., Shepherd, F. A. & Tsao, M. S. EGFR Mutations and Lung Cancer. *Annu. Rev. Pathol. Mech. Dis.* **6**, 49–69 (2011).

39. Midha, A., Dearden, S. & McCormack, R. EGFR mutation incidence in non-small-cell lung cancer of adenocarcinoma histology: a systematic review and global map by ethnicity (mutMapII). *Am. J. Cancer Res.* **5**, 2892–911 (2015).

40. Maheswaran, S. *et al.* Detection of Mutations in *EGFR* in Circulating Lung-Cancer Cells. *N. Engl. J. Med.* **359**, 366–377 (2008).

41. Seki, Y. *et al.* Picoliter-Droplet Digital Polymerase Chain Reaction-Based Analysis of Cell-Free Plasma DNA to Assess EGFR Mutations in Lung Adenocarcinoma That Confer Resistance to Tyrosine-Kinase Inhibitors. *Oncologist* **21**, 156–164 (2016).

42. de Biase, D. *et al.* Next-Generation Sequencing of Lung Cancer EGFR Exons 18-21 Allows Effective Molecular Diagnosis of Small Routine Samples (Cytology and Biopsy). *PLoS One* **8**, e83607 (2013).

43. Malapelle, U. *et al.* Profile of the Roche cobas® EGFR mutation test v2 for non-small cell lung cancer. *Expert Rev. Mol. Diagn.* **17**, 209–215 (2017).

44. Siegelin, M. D. & Borczuk, A. C. Epidermal growth factor receptor mutations in lung adenocarcinoma. *Lab. Investig.* **94**, 129–137 (2014).

45. Cole, R. H., Gartner, Z. J. & Abate, A. R. Multicolor Fluorescence Detection for Droplet Microfluidics Using Optical Fibers. *J. Vis. Exp.* e54010–e54010 (2016). doi:10.3791/54010

46. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).

47. Landau, D. A., Carter, S. L., Getz, G. & Wu, C. J. Clonal evolution in hematological malignancies and therapeutic implications. *Leukemia* **28**, 34–43 (2014).

48. Patel, J. P. *et al.* Prognostic Relevance of Integrated Genetic Profiling in Acute Myeloid Leukemia. *N. Engl. J. Med.* **366**, 1079–1089 (2012).

49. Buckley, S. A. & Walter, R. B. Antigen-specific immunotherapies for acute myeloid leukemia. *Hematology* **2015**, 584–595 (2015).

50. García-Dabrio, M. C. *et al.* Complex Measurements May Be Required to Establish the Prognostic Impact of Immunophenotypic Markers in AML. *Am. J. Clin. Pathol.* **144**, 484–492 (2015).

51. Papaemmanuil, E. *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N. Engl. J. Med.* **374**, 2209–2221 (2016).

52. Paguirigan, A. L. *et al.* Single-cell genotyping demonstrates complex clonal diversity in acute myeloid leukemia. *Sci. Transl. Med.* **7**, 281re2 (2015).

53. Wang, L. *et al.* Integrated single-cell genetic and transcriptional analysis suggests novel drivers of chronic lymphocytic leukemia. *Genome Res.* **27**, 1300–1311 (2017).

54. Pellegrino, M. *et al.* High-throughput single-cell DNA sequencing of acute myeloid leukemia tumors with droplet microfluidics. *Genome Res.* **28**, 1345–1352 (2018).

55. Smith, C. C. *et al.* Heterogeneous resistance to quizartinib in acute myeloid leukemia revealed by single-cell analysis. *Blood* **130**, 48–58 (2017).

56. De Zen, L. *et al.* Quantitative multiparametric immunophenotyping in acute lymphoblastic leukemia: correlation with specific genotype. I. ETV6/AML1 ALLs identification. *Leukemia* **14**, 1225–1231 (2000).

57. van Galen, P. *et al.* Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. *Cell* **0**, 1–17 (2019).

58. Suvà, M. L. & Tirosh, I. Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges. *Mol. Cell* **75**, 7–12 (2019).

59. Giustacchini, A. *et al.* Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. *Nat. Med.* **23**, 692–702 (2017).

60. Nam, A. S. *et al.* Somatic mutations and cell identity linked by Genotyping of Transcriptomes. *Nature* **571**, 355–360 (2019).

61. Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).

62. Zeisel, A. *et al.* Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–42 (2015).

63. Schiebinger, G. *et al.* Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell* **176**, 928-943.e22 (2019).

64. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science (80-.).* **343**, 193–196 (2014).

65. Mansour, M. R. *et al.* Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 1373–7 (2014).

66. Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102–111 (2020).

67. Shahi, P., Kim, S. C., Haliburton, J. R., Gartner, Z. J. & Abate, A. R. Abseq: Ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding. *Sci. Rep.* **7**, 1–12 (2017).

68. Schuurhuis, G. J. *et al.* Minimal/measurable residual disease in AML: a consensus document from the European LeukemiaNet MRD Working Party. *Blood* **131**, 1275–1291 (2018).

69. Wood, B. L. Flow Cytometric Monitoring of Residual Disease in Acute Leukemia. in 123–136 (Humana Press, Totowa, NJ, 2013). doi:10.1007/978-1-62703-357-2_8

70. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).

71. Appelbaum, F. R. & Bernstein, I. D. Gemtuzumab ozogamicin for acute myeloid leukemia. *Blood* **130**, 2373–2376 (2017).

72. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (2018).

73. Campana, D. & Behm, F. G. Immunophenotyping of leukemia. *J. Immunol. Methods* **243**, 59–75 (2000).

74. Sexauer, A. *et al.* Terminal myeloid differentiation in vivo is induced by FLT3 inhibition in FLT3/ITDAML. *Blood* **120**, 4205–4214 (2012).

75. McMahon, C. M. *et al.* Gilteritinib induces differentiation in relapsed and refractory FLT3-mutated acute myeloid leukemia. *Blood Adv.* **3**, 1581–1585 (2019).

76. Yun, H. D. *et al.* Erythroid differentiation of myeloblast induced by gilteritinib in relapsed FLT3-ITD–positive acute myeloid leukemia. *Blood Advances* **3**, 3709–3712 (2019).

77. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).

78. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. (2008). doi:10.1088/1742-5468/2008/10/P10008

79. Buscarlet, M. *et al.* DNMT3A and TET2 dominate clonal hematopoiesis and demonstrate benign phenotypes and different genetic predispositions. *Blood* **130**, 753–762 (2017).

80. Jaiswal, S. *et al.* Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *N. Engl. J. Med.* **377**, 111–121 (2017).

81. Gong, H. *et al.* Simple Method To Prepare Oligonucleotide-Conjugated Antibodies and Its Application in Multiplex Protein Detection in Single Cells. *Bioconjug. Chem.* **27**, 217–225

(2016).

82. Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S. & Marioni, J. C. Normalizing single-cell RNA sequencing data: Challenges and opportunities. *Nature Methods* **14**, 565–571 (2017).

83. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Benjamin Demaree
BD992956C1574DE...
Author Signature

5/11/2020
Date